ALCALDIA MAYOR
DE BOGOTA D.C.
Secretaría de
Educación

v 1.2
3/14/2019

Arquitectura Integración datos

Secretaria de Educación Distrital de Bogotá

softwareONE
The Software Licensing Experts™

Control de cambios

N °	Fecha	Autor	Modificación	Referencia del Cambio
1	05/02/2019	Carlos Ardila	Solicitud de ajustes	Versión 1.0
2	14/02/2019	Carlos Ardila	Solicitud de ajustes	Versión 1.0
3	01/03/2019	Carolina Gómez	Ajustes realizados	Versión 1.1
4	14/03/2019	Carolina Gómez	Ajustes realizados	Versión 1.2

189

Tabla de contenido

1. Documento dirigido a: **6**
2. Marco de desarrollo **6**
2.1. Descripción del problema 6
2.2. Justificación 7
3. Objetivos **7**
3.1. Objetivo general 7
3.2. Objetivos específicos 8
4. Caracteristicas de la plataforma **8**
4.1. Servicios de análisis de Azure 8
4.2. Bases de Datos 10
4.3. Data Factory 12
4.4. DataBase 16
4.5. Vista de Datos 17
4.6. PowerBI 18
5. Analisis de Arquitectura Inicial **18**
5.1. Información 18
5.2. Sistemas de información 19
5.3. Administración de datos 20
5.4. Estructura de datos SIMAT 20
5.5. Estructura de datos Acceso y permanencia 21
5.6. Estructura de datos Escalafón 22
5.7. Estructura de datos Servicios Públicos 23
5.8. Estructura de datos Instituciones 24
6. Diseño Arquitectura Nueva **24**
6.1. Arquitectura de la solución 24
6.2. Componentes 26
6.3. Ventajas para el negocio 28
6.4. Sistemas de información 28
6.5. Por qué un Repositorio de datos, características evaluadas: 29

199

6.6. Administración de Datos 30
6.7. Vistas de datos 31
6.7.1. Vista de Datos Docentes 32
6.7.2. Vista de datos Estudiantes 34
6.7.3. Vista de datos instituciones 36
6.7.4. Relación de vista de datos en el Repositorio Datos 40
6.7.5. Ejecución de procesos por Vista de datos 40
7. Desarrollo del proyecto 41
7.1. Integración Fuente Simat 41
7.1.1. Definición de la fuente: 42
7.1.2. Procesos: 42
7.1.2.1. Extracción 43
7.1.2.2. Transformación: 43
7.1.2.3. Carga: 44
7.1.2.4. Programación de ejecución de procesos 44
7.2. Integración Fuente Simat Detallado. 45
7.2.1. Definición de la fuente. 45
7.2.1.1. Procesos: 46
7.2.1.2. Extracción: 46
7.2.1.3. Transformación: 47
7.2.1.4. Carga: 48
7.2.1.5. Programación de ejecución de procesos 48
7.3. Integración Fuente DUEB. 48
7.3.1. Definición de la fuente: 49
7.3.2. Procesos: 49
7.3.2.1. Extracción: 49
7.3.2.2. Carga: 50
7.3.2.3. Programación de ejecución de procesos 51
7.4.1. Definición de la fuente. 51
7.4.1.1. Extracción: 51
7.4.1.2. Transformación: 54
7.4.1.3. Carga: 55

7.4.1.4. Programación de ejecución de procesos 55
7.5. Integración Fuente Humano. 56
7.5.1. Definición de la fuente: 56
7.5.2. Procesos: 56
7.5.2.1. Extracción 57
7.5.2.2. Transformación: 57
7.5.2.3. Carga: 58
7.5.2.4. Programación de ejecución de procesos 58
7.6. Indicadores 59
8. Recomendaciones 60

192

1. Documento dirigido a:

Este documento va dirigido a profesional en ingeniería de sistemas con conocimientos específicos en diseño, bases de datos, análisis de datos y herramientas de integración de información.

2. Marco de desarrollo

2.1. Descripción del problema

La dirección de cobertura que reporta a la Subsecretaría de acceso y permanencia es la encargada de la transformación del archivo de matrícula mediante la descarga mensual del anexo 6ª del sistema de matrículas ("SIMAT") provisto por el Ministerio de Educación y que se carga en Dynamics 365 para obtener la vista 360 grados del niño a partir de la información cargada por los archivos generados en el SIMAT. El proceso de carga de la información se realiza a través de un integrador (SMART 360) de datos propietario de un desarrollador de software tipo ISV[1] (Independent Software Vendor).

El área técnica de la secretaria realiza pruebas de carga de la información a Dynamics 365 encontrando que la herramienta realiza la tarea en espacios de tiempo de aproximadamente 3 a 4 días sin realizar las tareas de calidad de datos y aun procesando los archivos de simat a través del área, adicionalmente ocupando un espacio importante en la base de datos de Dynamics 365 cuyo precio por giga es elevado versus el almacenamiento en base de datos en modelos PAAS[2] o IAAS[3].

[1] https://en.wikipedia.org/wiki/Independent_software_vendor

[2] https://crearsoftware.com/2008/08/14/definicion-de-paas/

[3] https://www.rackspace.com/es-co/library/what-is-iaas

193

Adicionalmente existen otras fuentes como humano y los procesos desarrollados en Dynamics 365 con información sobre los procesos de acceso y permanencia, escalafón de docentes privados y la necesidad creciente de la secretaria de integrar nuevas fuentes de datos para enriquecer la información de su vista 360 grados.

2.2. Justificación

Las áreas usuarias de la entidad adolecen de una herramienta que les provea información única para la ejecución de sus procesos. Una sola versión de la información se hace imperiosa frente al desafío de la toma de decisiones.

La toma de decisiones se basa en seleccionar las estrategias indicadas de acuerdo a la información existente en la entidad, la precisión de la misma juega un papel fundamental en la elaboración y ejecución de las iniciativas que llevarán a un nuevo ciclo de inmersión, análisis de los resultados obtenidos y permitirá avanzar en aras de la mejora administrativa y educativa del distrito.

La arquitectura de esta solución busca apoyar el rendimiento en las operaciones de las áreas que hacen uso de la información base para la Secretaría de Educación del Distrito y la toma de decisiones con los datos históricos y actuales.

3. Objetivos

3.1. Objetivo general

Integrar las fuentes de datos para las áreas funcionales de la Secretaría mediante tecnologías inclusivas y de alto desempeño.

3.2. Objetivos específicos

Diseñar una solución de integración de datos que permita consolidar la información de las fuentes: SIMAT, Humano, los procesos en Dynamics 365 e incluir nuevas fuentes de datos en el futuro.

Diseñar un almacén de datos que le permita a la SED contar con la vista 360 de niños, docentes e instituciones.

4. Caracteristicas de la plataforma

Dentro del marco del contrato **CO1.PCCNTR.385761** de licenciamiento realizado con Microsoft se cuenta con la suscripción de Microsoft Azure que entrega una plataforma con tecnologías para el análisis de datos en versión PAAS[4] e IAAS[5] que le permitirán a la secretaría explotar todos los beneficios de la nube con modelos flexibles y reduciendo el impacto en la gestión de dichos servicios logrando eficiencia administrativa.

4.1. Servicios de análisis de Azure[6]

Es un conjunto integral de soluciones que convierten los datos en información práctica.

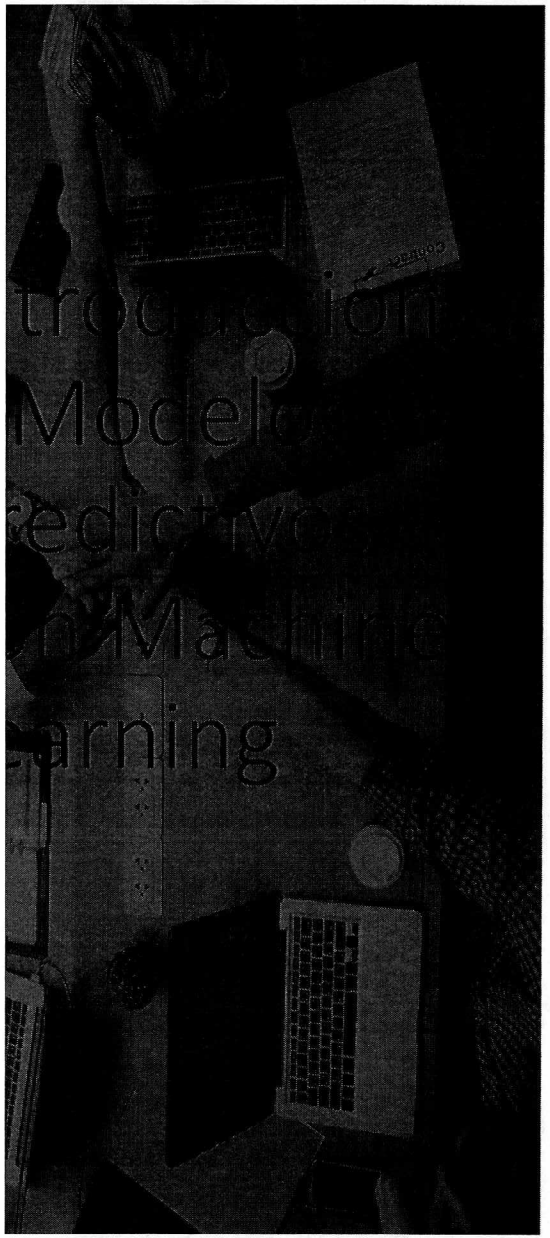
Permitiendo obtener el máximo valor de las llamadas de macrodatos de soluciones de análisis que vayan más allá de las funcionalidades de las aplicaciones de procesamiento de datos tradicionales. Azure ofrece un conjunto integral de soluciones de macrodatos que ayudan a recopilar,

[4] https://azure.microsoft.com/es-es/overview/what-is-paas/

[5] https://azure.microsoft.com/es-es/overview/what-is-iaas/

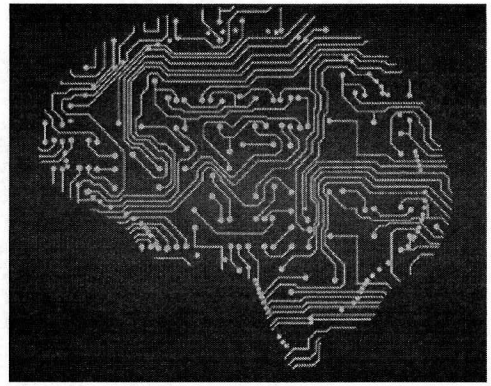
[6] https://azure.microsoft.com/es-es/product-categories/analytics/


Agenda
1. Introducción
2. Arquitectura
3. Diagrama de flujo
4. Presentación del modelo
4.1. Análisis Exploratorio de Datos
4.2. Análisis de variables
4.3. Modelos predictivos
4.4. Resultados
software ONE



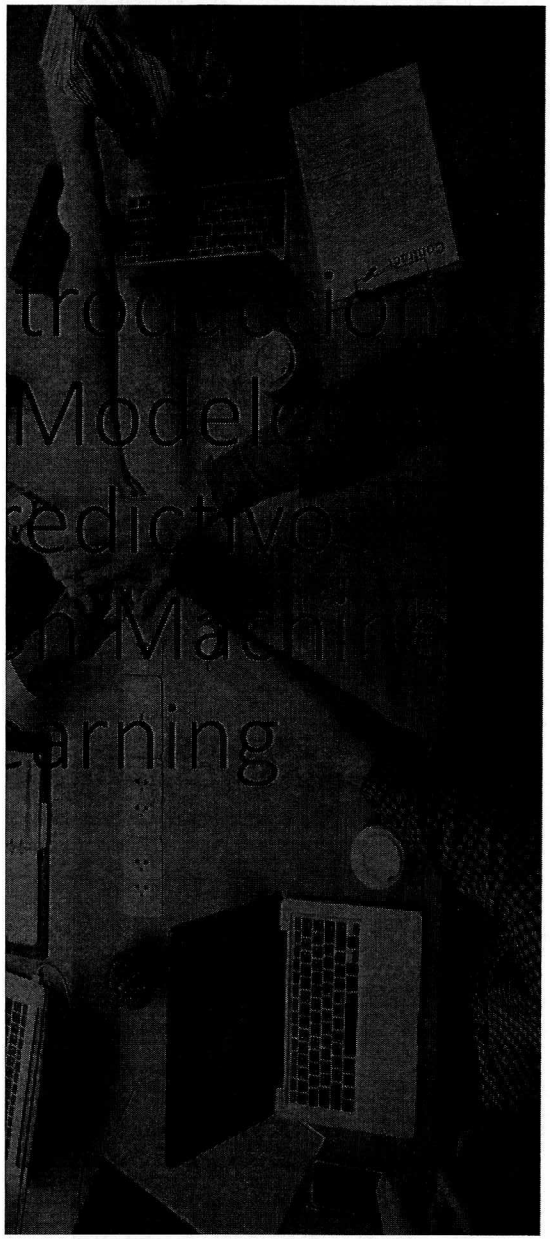
Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos partiendo de datos y ser capaz de predecir comportamientos futuros.



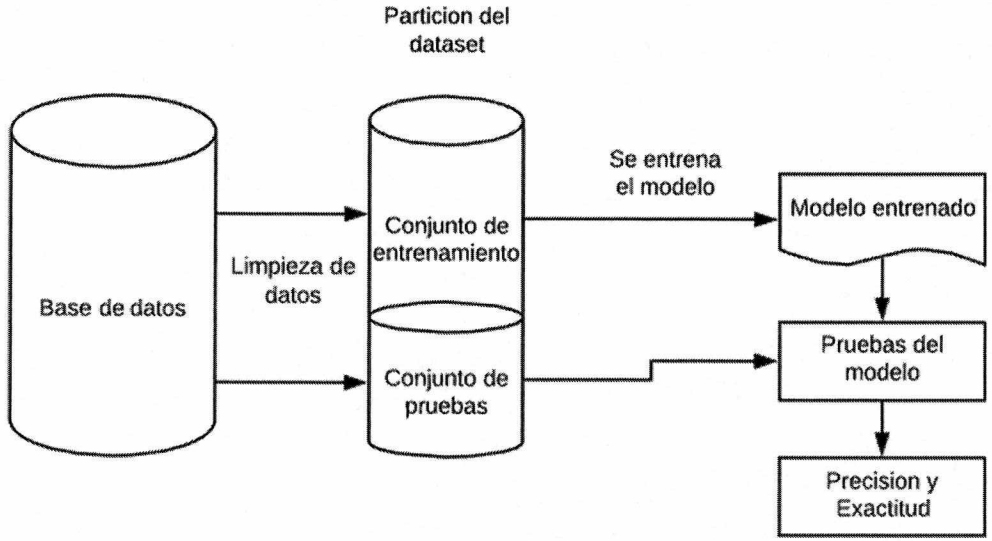


Tomado de: CleverData.io

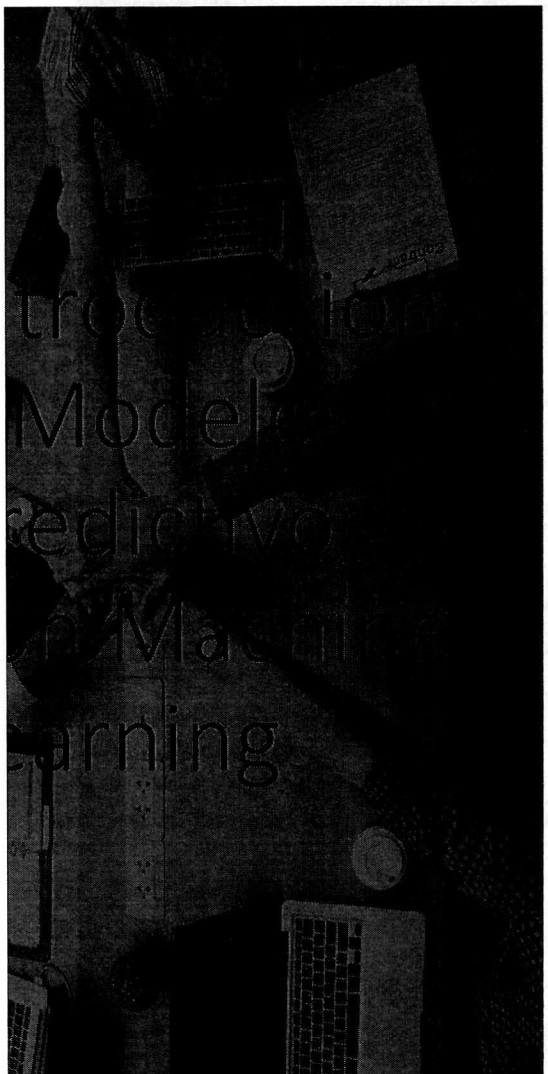
softwareONE



Con la finalidad de que un modelo aprenda, es necesario entrenarlo, para ello el dataset (conjunto de datos) original es divido en dos: entrenamiento y prueba. Si los resultados no son óptimos debe realizarse cambios en el entrenamiento (suponiendo que los datos son de calidad).

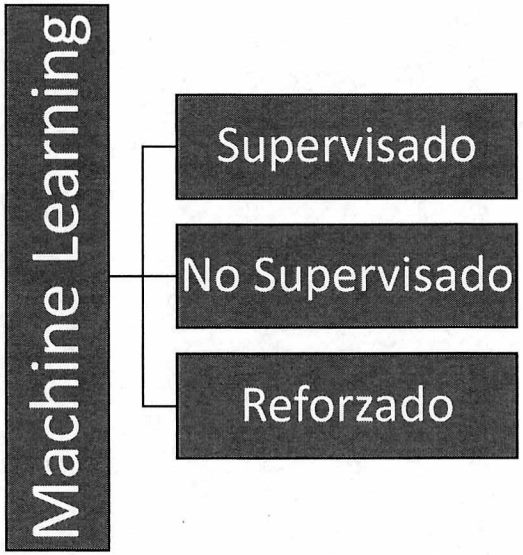


softwareONE

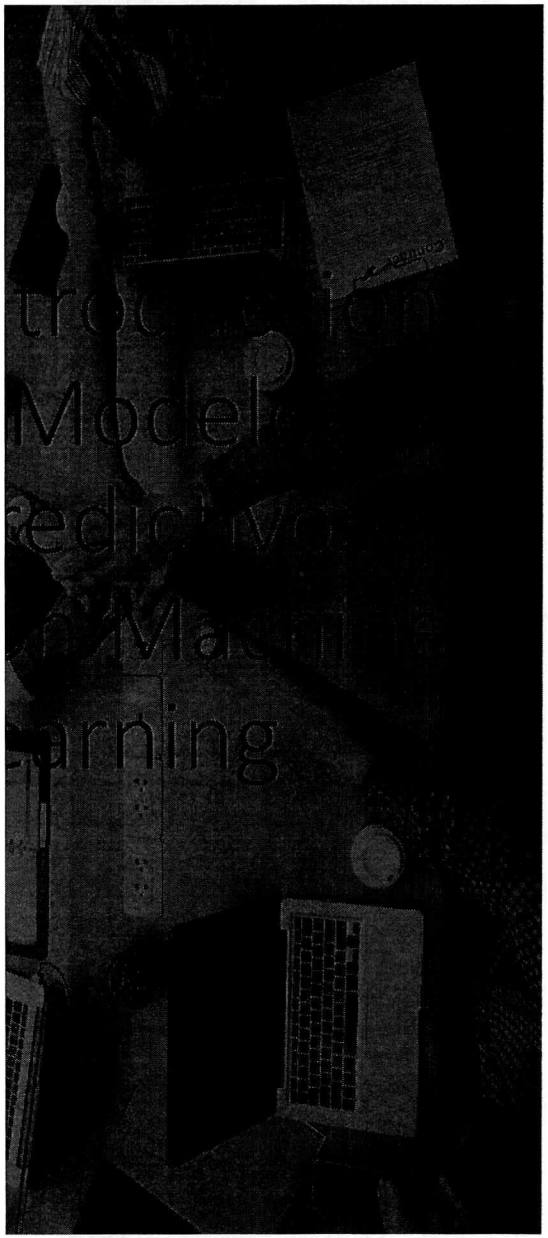


Los modelos de Machine Learning están clasificados en:

- Supervisado
- No supervisado
- Reforzado

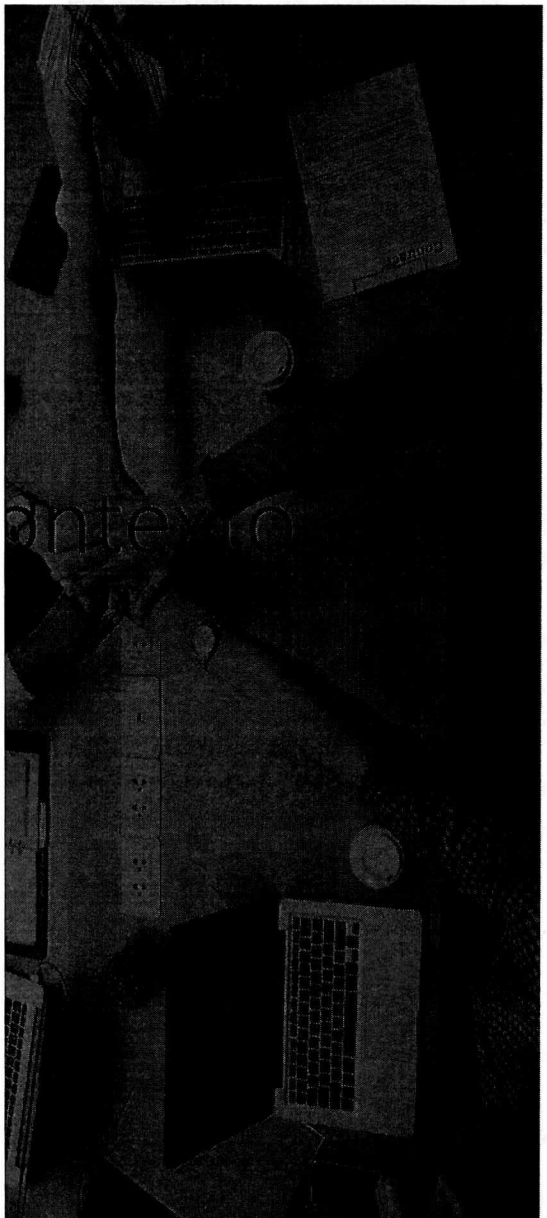


softwareONE

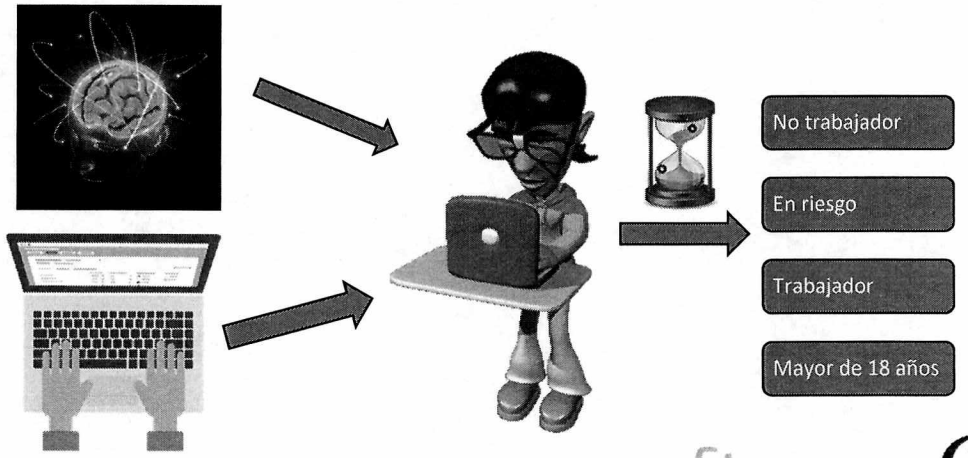


Cuando se analiza información proveniente de una base de datos en el ámbito de Machine Learning, es importante conocer que las columnas de la tabla son más conocidas como variables y las filas como registros o instancias. En el caso de los modelos supervisados existen dos tipos de variables: una variable objetivo y una o muchas variables predictoras.

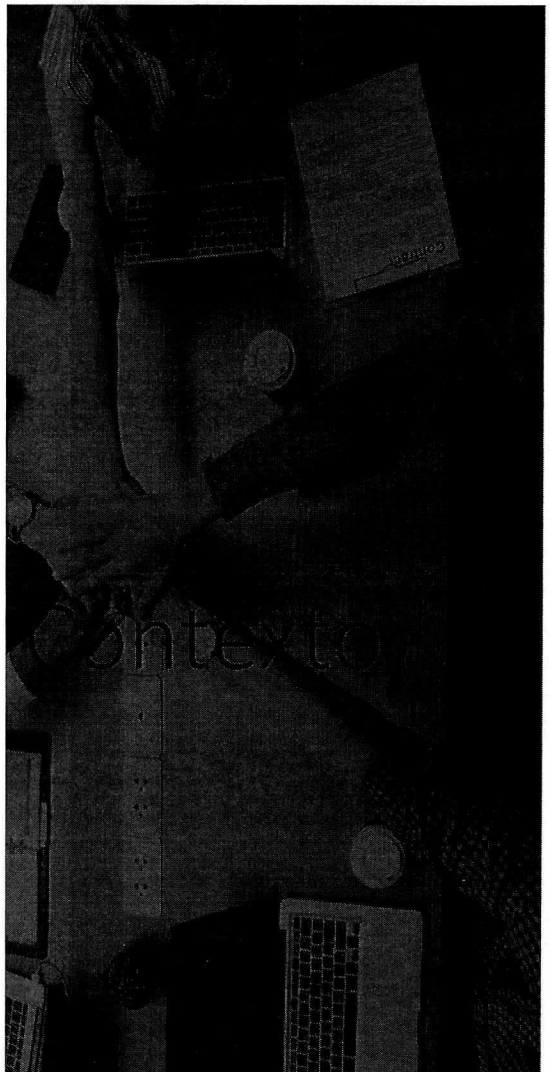
softwareONE



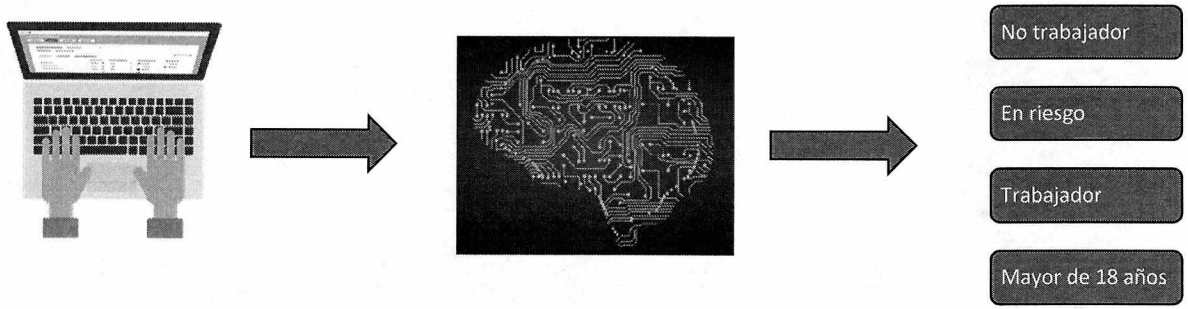
Actualmente en la Secretaría de Educación del Distrito existe un área encargada de realizar un proceso de categorización del riesgo laboral en el que se encuentran los estudiantes. Para lograr dicha categorización, se analiza una encuesta que es diligenciada por los estudiantes en la cual se les preguntan determinados aspectos socio-demográficos, familiares y académicos. Adicionalmente, en dicha encuesta deben diligenciar una pregunta abierta que corresponde a "*Actividad en el Tiempo Libre*", donde los estudiantes pueden escribir con sus propias palabras las actividades que realizan cuando están fuera de las instituciones educativas. Este proceso de identificación de estudiantes menores de edad que se encuentran en riesgo de trabajo infantil es realizado por profesionales (en su mayoría psicólogos) que analizan la información de la encuesta; es un procedimiento realizado manualmente, con unas ciertas 'reglas' que los profesionales emplean.



softwareONE



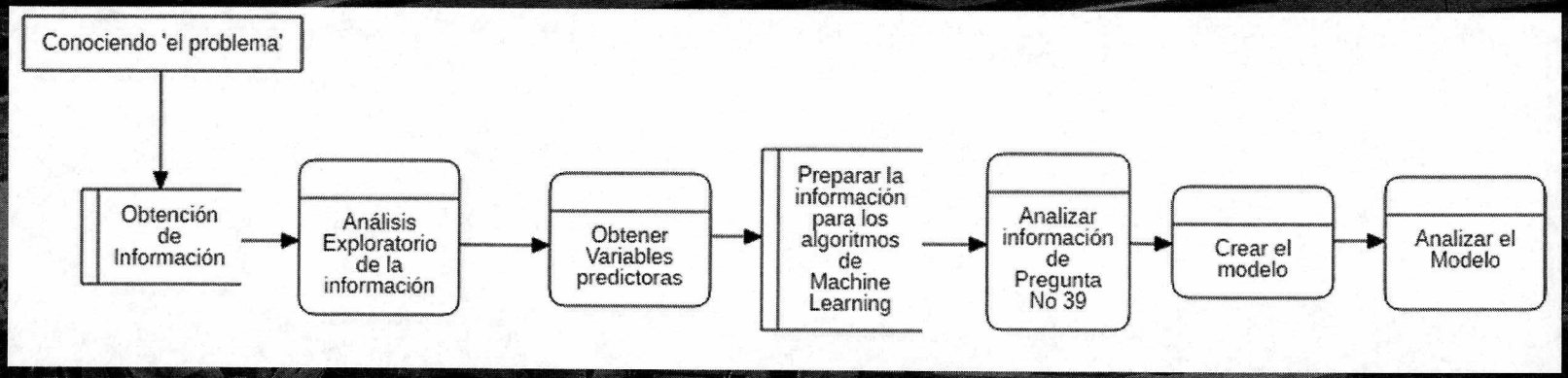
Se presenta una solución basados en algoritmos de Machine Learning clasificados y Procesamiento de Lenguaje Natural con la finalidad de lograr clasificar estos estudiantes de una manera más rápida, eficiente, menos propensa a errores y desde luego más económica para la Secretaría de Educación. Sin embargo, se debe aclarar que **tanto el modelo analítico y de clasificación, como los resultados obtenidos por el mismo, reflejan la precisión con la que los profesionales han realizado su trabajo, puesto que el modelo es entrenado mediante estos datos.**



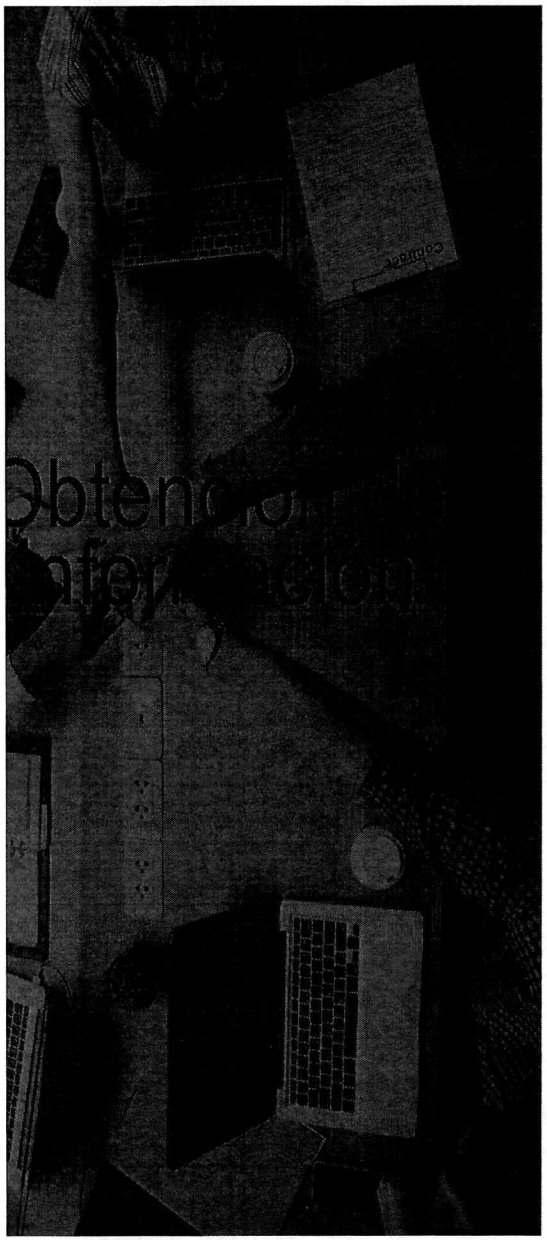
softwareONE

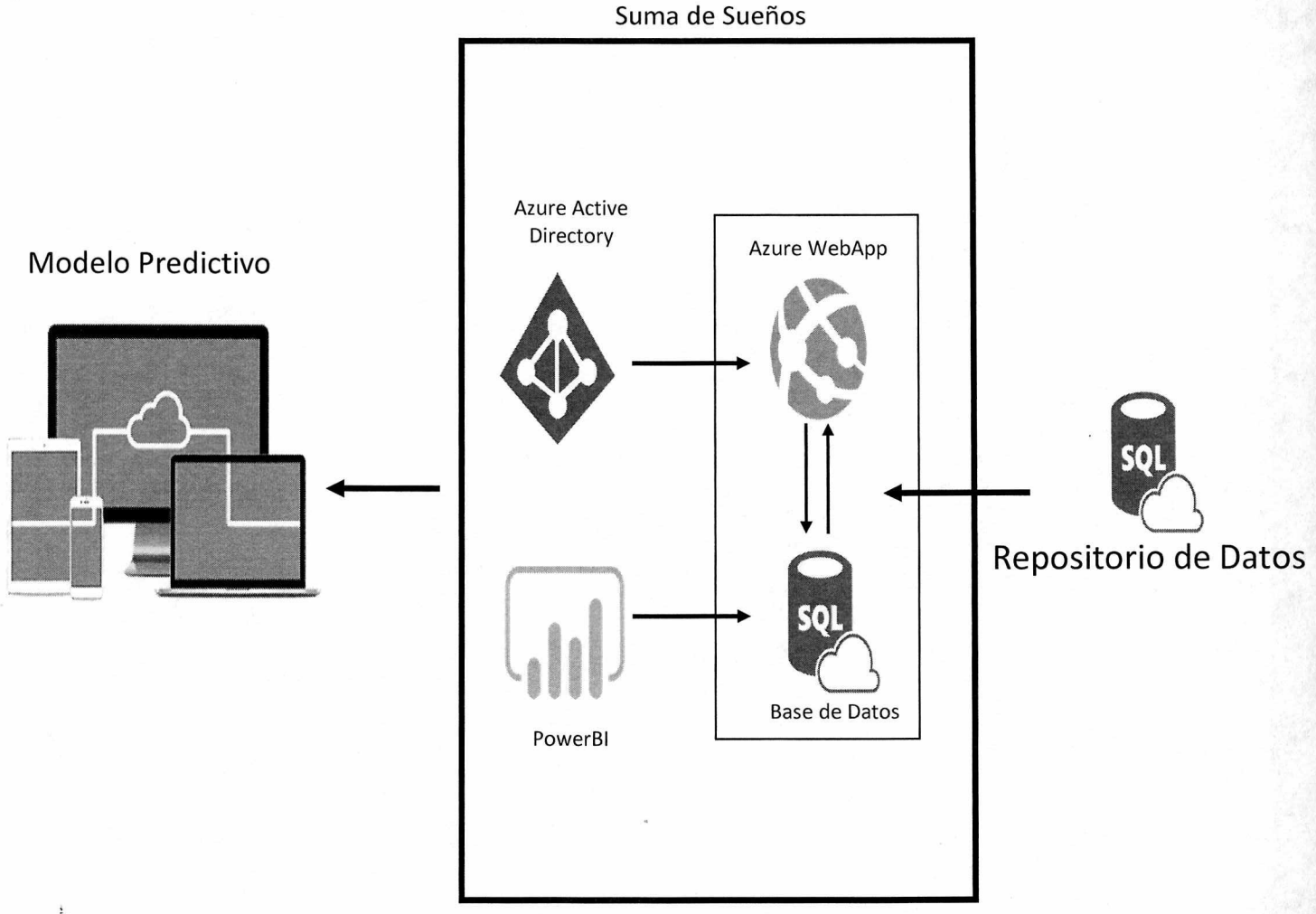
202

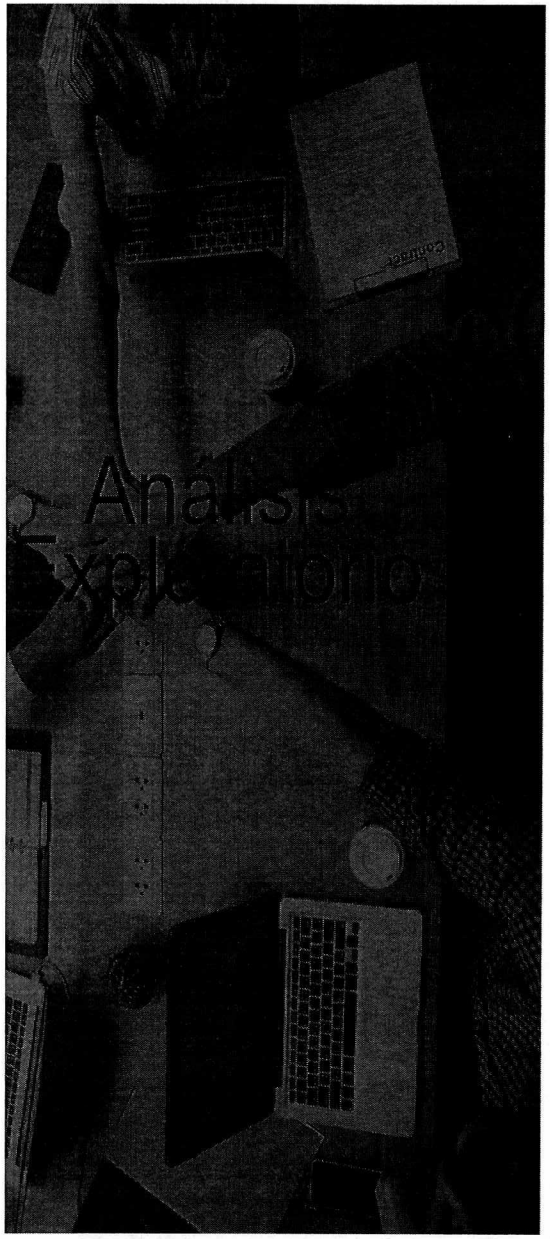
Diagrama de Flujo



softwareONE



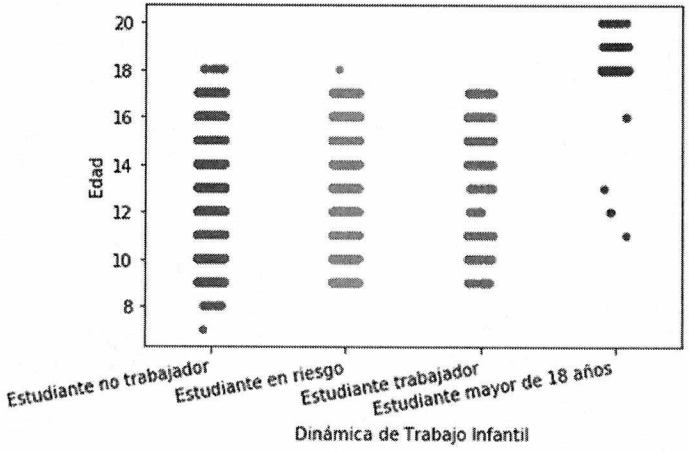

Suma de Sueños
Azure Active Directory
Azure WebApp
Modelo Predictivo
SQL
Repositorio de Datos
SQL
PowerBI
Base de Datos

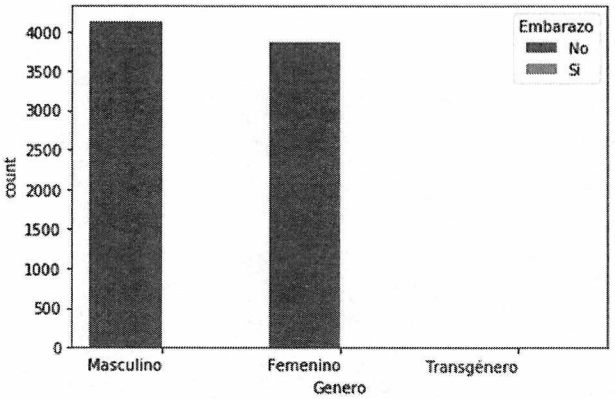


ANÁLISIS DE UNA VARIABLE

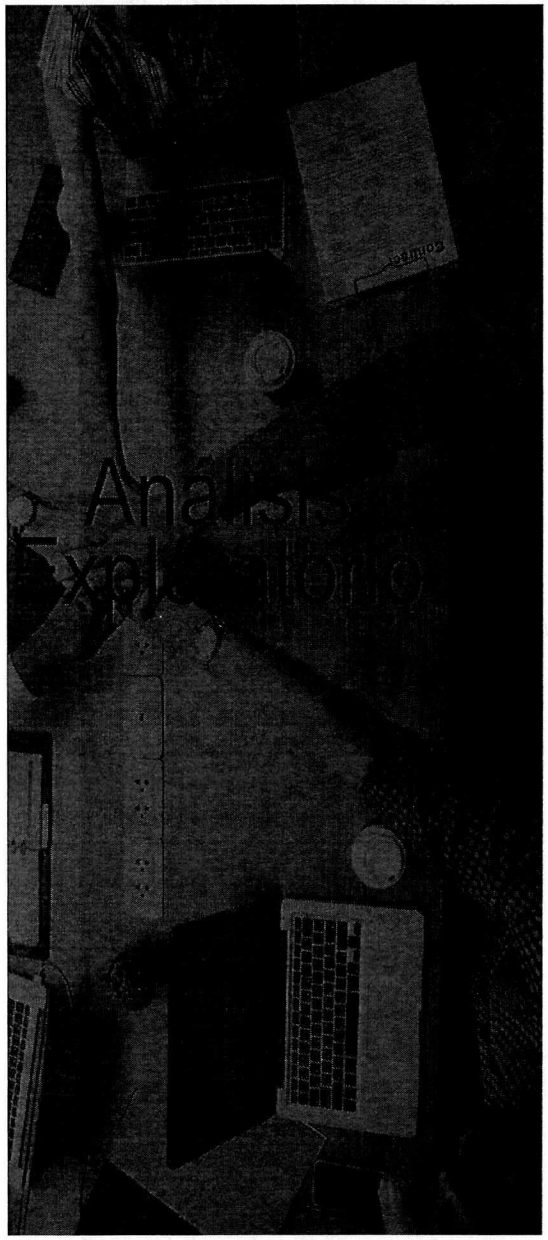
La finalidad de esta fase es obtener información sobre los datos que permitan realizar conjeturas sobre los datos; también se evalúa la calidad de los mismos y se conoce mejor el problema desde la perspectiva de datos.
Se eliminan registros inconsistentes o poco lógicos.
En esta fase se eliminan variables predictoras que estadísticamente aportan muy poca información a la variable objetivo.

Se muestra de manera breve algunas inconsistencias del dataset:

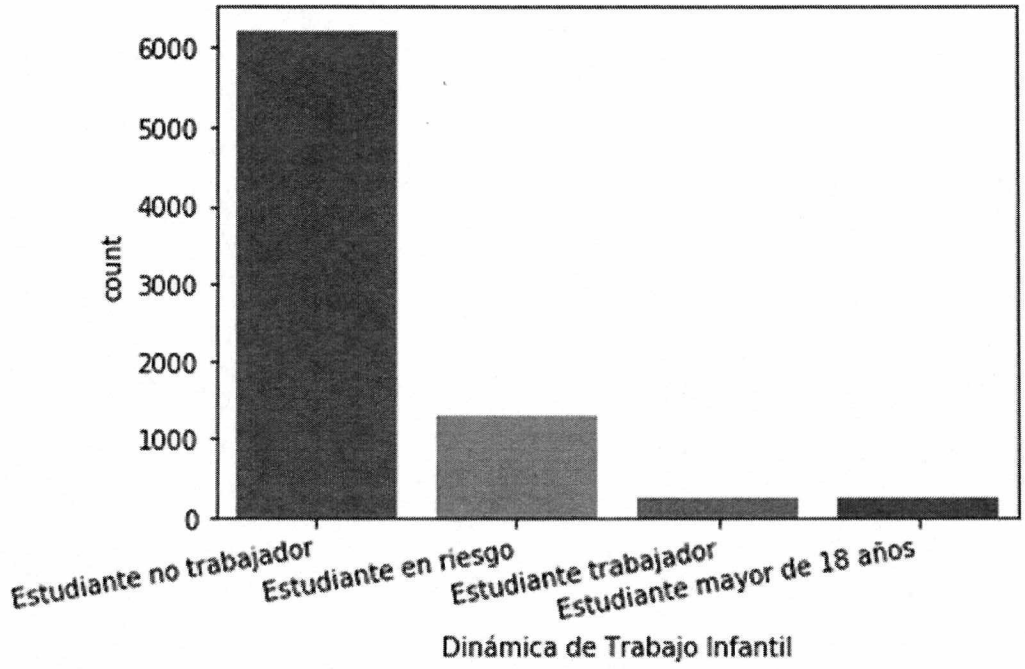




Genero	Embarazo
Masculino	Si
Masculino	Si
Masculino	Si

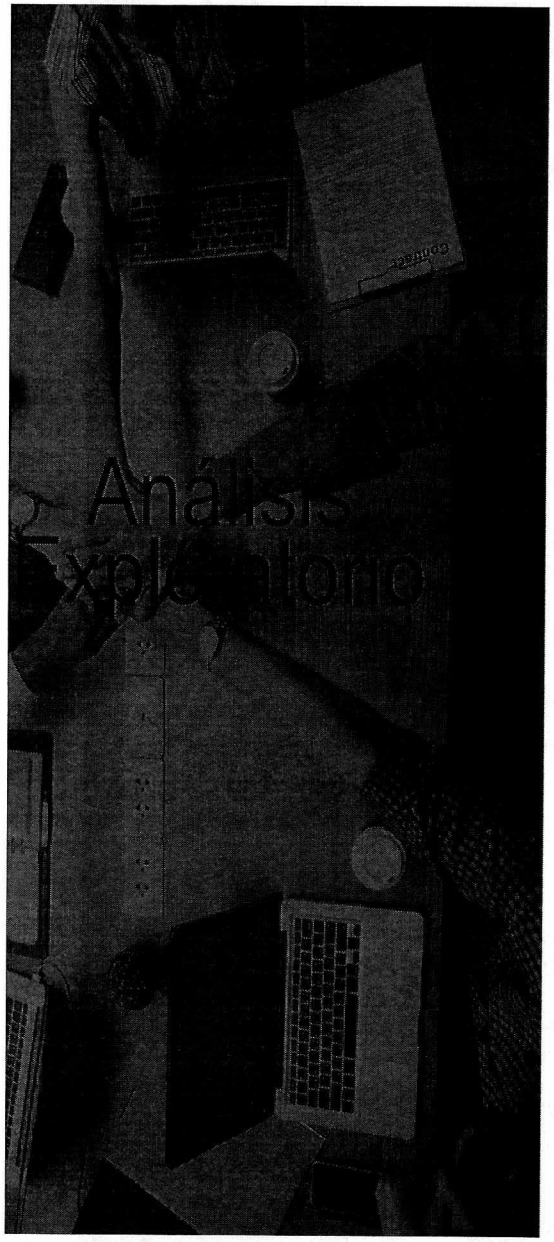


COMPORTAMIENTO DE LA VARIABLE OBJETIVO




```
#####Porcentaje#####
Estudiante no trabajador        0.774535
Estudiante en riesgo            0.159995
Estudiante trabajador           0.033171
Estudiante mayor de 18 años     0.032298
```

softwareONE



ANÁLISIS DE DOS VARIABLES

En esta sección se utilizan tablas de contingencia (que está automatizadas en el documento técnico) para obtener información relevante entre dos variables categóricas.

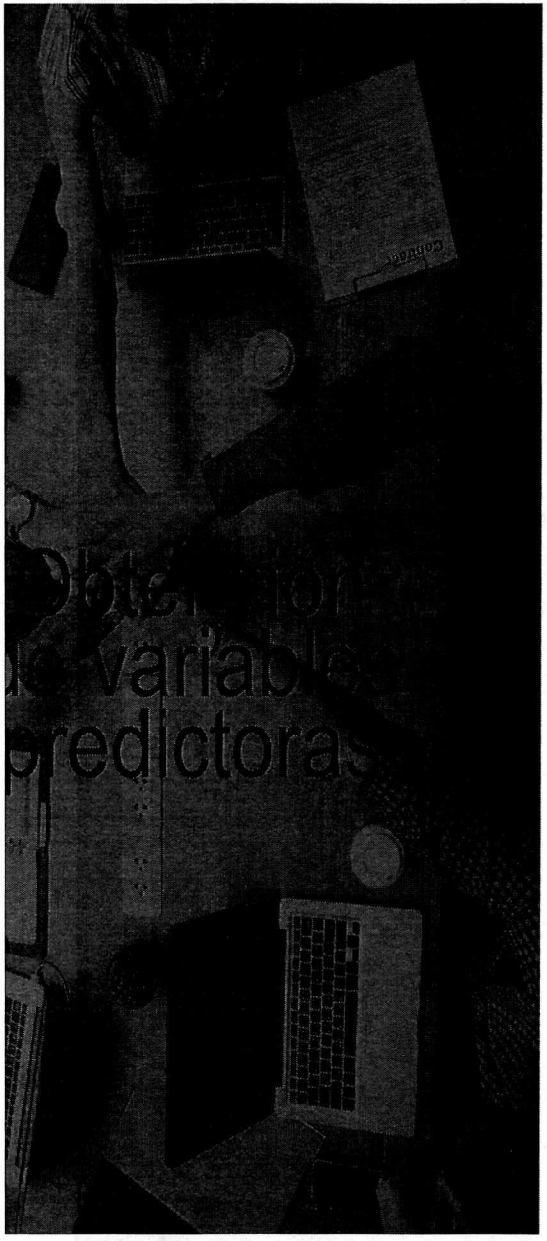
Debido a que no es posible mostrar una tabla de contingencia resultado para el caso específico, se muestra un ejemplo sencillo sobre el uso de estas tablas.

En el ejemplo se tiene dos variables categóricas una con valores (A, B, C) y otra con valores (Bajo, Medio, Alto). Gracias al uso de las variables de contingencia se puede 'numerizar' la información categórica y de esta manera encontrar, relevancia en la información

	Bajo	Medio	Alto	Total
A	14	22	38	**74**
B	10	11	49	**70**
C	6	11	48	**65**
Total	**30**	**44**	**135**	**209**

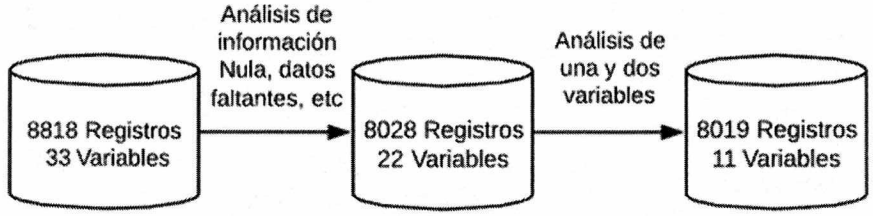
Ejemplo tabla de contingencia

softwareONE

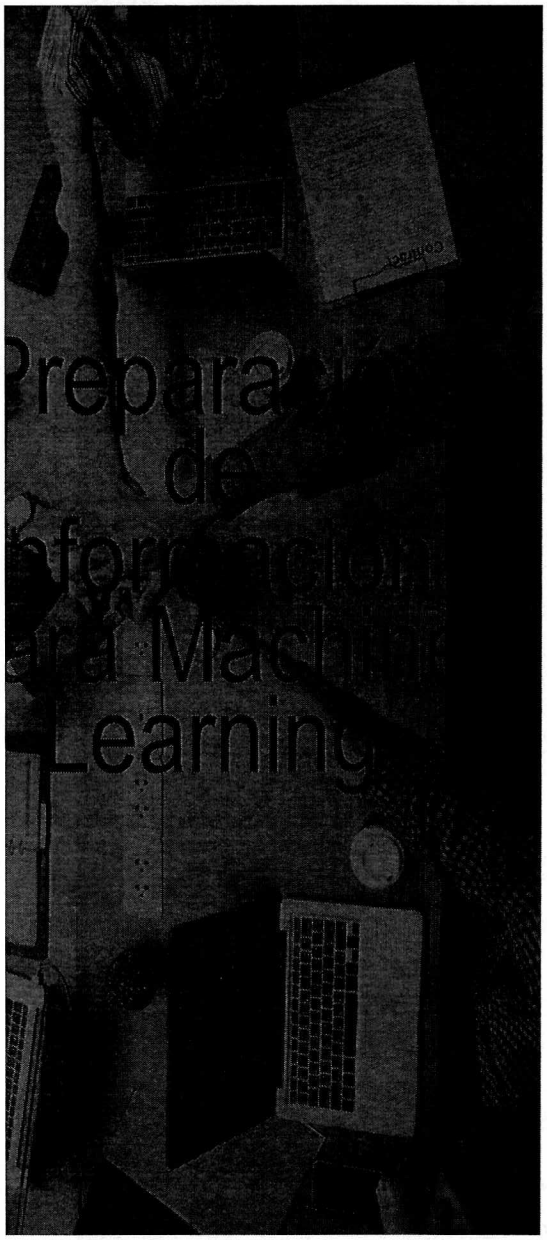


RESULTADOS DE LIMPIEZA

Después del proceso de limpieza, el dataset se reduce de 8818 registros con 33 variables a 8018 registros con 11 variables.



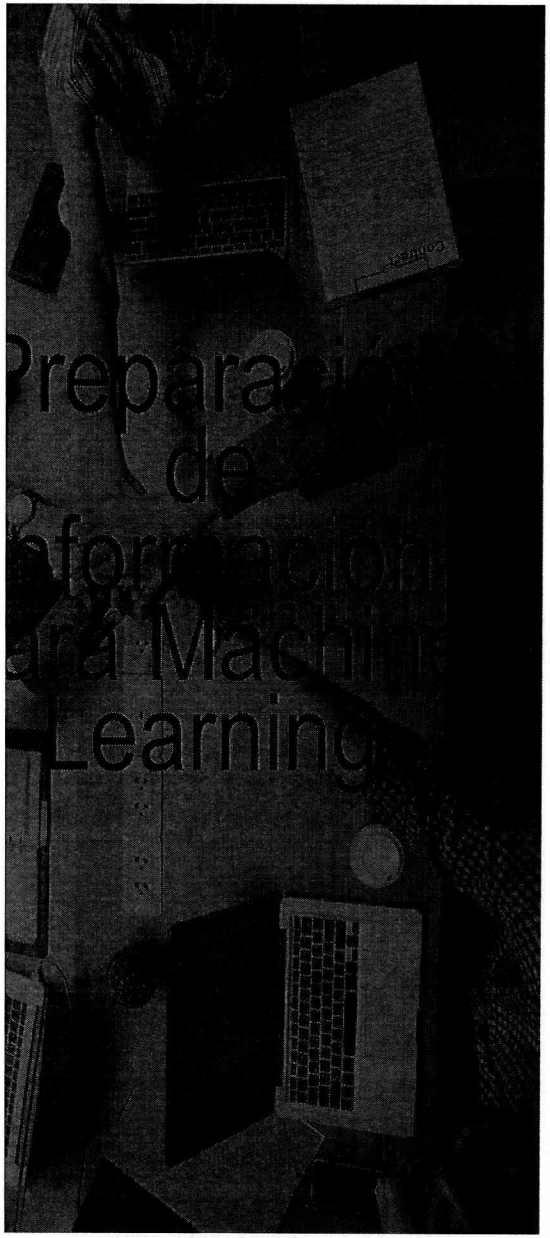
softwareONE



Debido a que la gran mayoría de algoritmos de Machine Learning trabajan con datos numéricos, se deben transformar los datos de texto (o categóricos) a número. Adicionalmente, se revisa la información numérica con la finalidad de que ésta también se encuentre en la misma escala de información para que los algoritmos se comporten mejor.



softwareONE

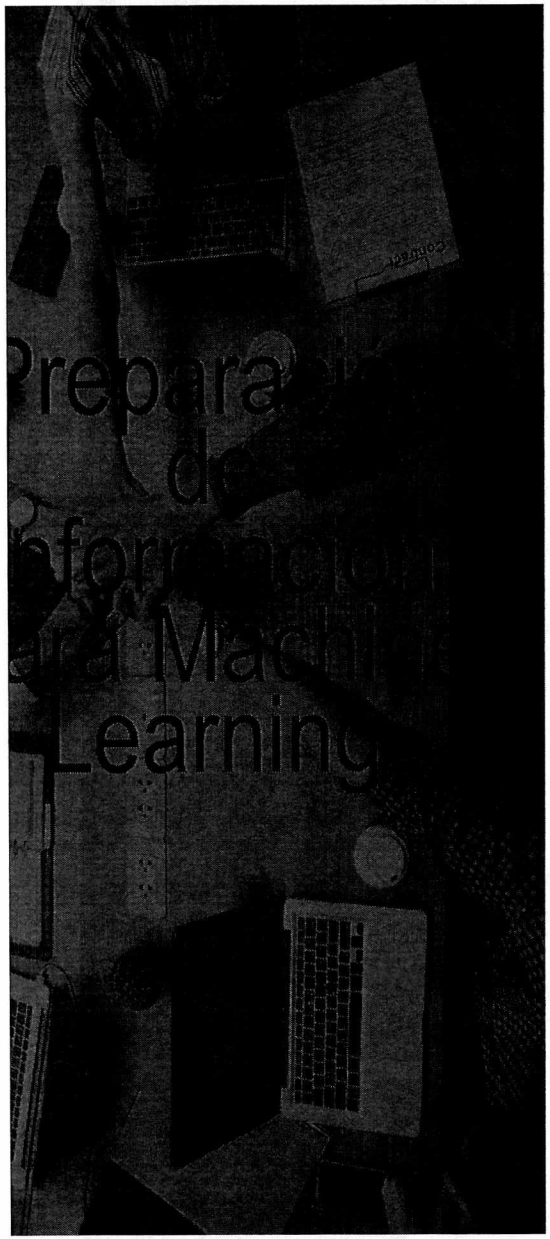


INFORMACIÓN CATEGÓRICA

El proceso que se realizó fue:

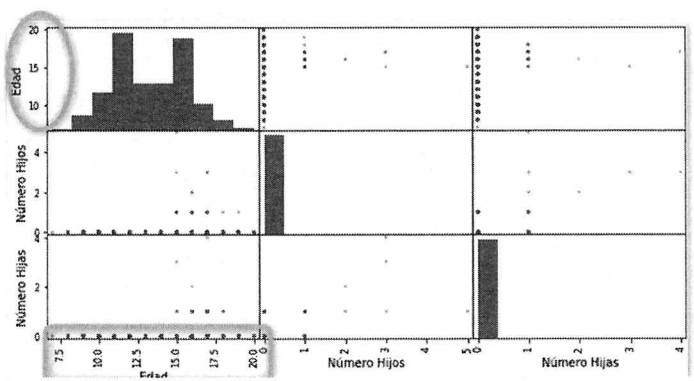
1. Separar las columnas categóricas del dataset al cual se le realizó el proceso de data cleaning sin tomar en cuenta tanto la variable objetivo 'Dinámica de Trabajo Infantil' como la variable a la cual se le realizará el Procesamiento Natural de Lenguaje 'Actividad Tiempo Libre',
2. Crear variables dummies de esta información (lo cual aumentará significativamente las columnas del dataset).

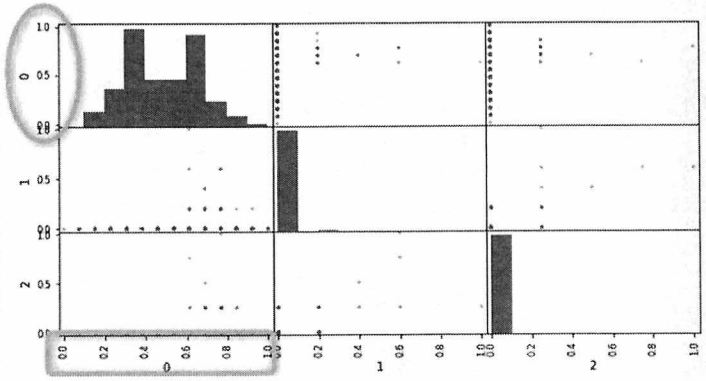
softwareONE



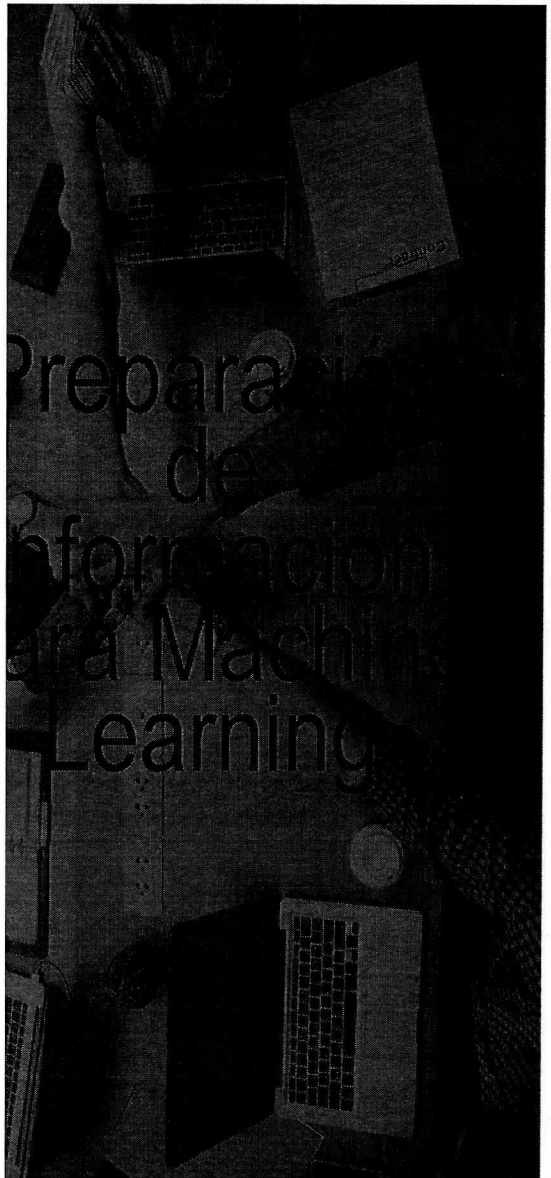
INFORMACIÓN NUMÉRICA

El proceso para preparar este tipo de información para ML será solo de 'Escalar' o 'Normalizar' la información.





softwareONE



PREGUNTA ABIERTA
'Actividad Tiempo Libre'

En procesamiento de lenguaje natural se utilizan una serie de técnicas en programación para eliminar o mejorar la forma en la cual se expresa la información. Para lograrlo se eliminan ciertas palabras como "y," "que", "la", entre otras, conocidas como 'stopwords'. También se eliminan signos de puntuación y se 'lematizan' las palabras, esto es, analizar cada palabra tomando en cuenta un análisis morfológico de las palabras; un ejemplo de esto podría ser: niñas, cuya información morfología es que es el plural de niños, y su *'lema'* sería niño.
También se realiza una bolsa de palabras, la cual es vectorizada y se calculan pesos.

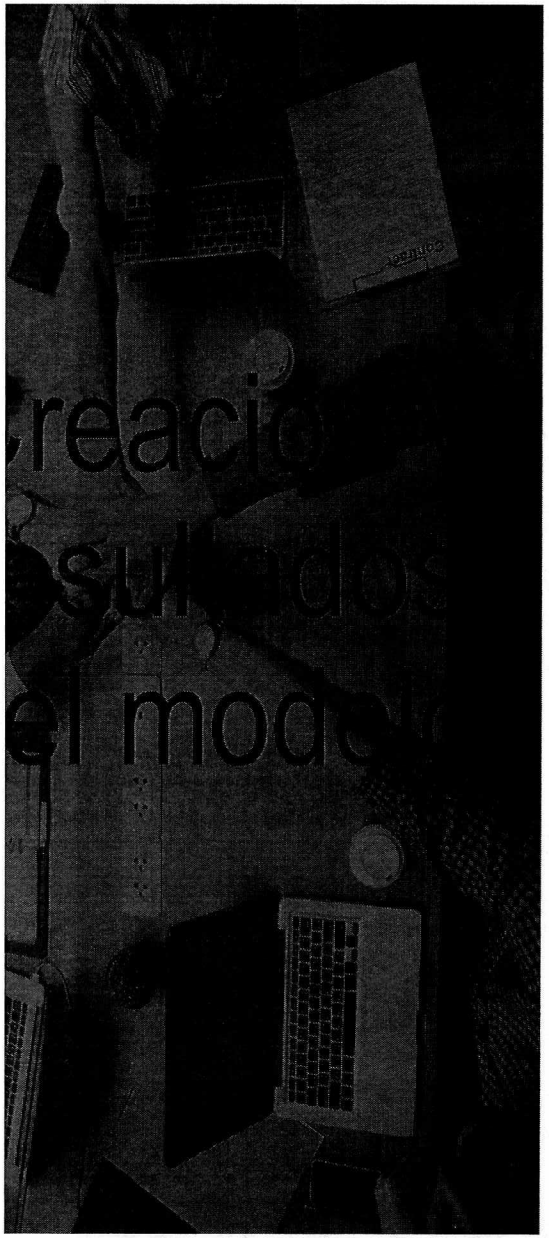
	and	another	cats	cheese	dogs	example	mouse	simple	with
0	0.0	0.000000	0.067578	0.000000	0.000000	0.0	0.067578	0.0	0.0
1	0.0	0.057924	0.057924	0.000000	0.156945	0.0	0.000000	0.0	0.0
2	0.0	0.057924	0.000000	0.156945	0.000000	0.0	0.057924	0.0	0.0

softwareONE

Al final del procedimiento de preparación de información se obtiene un dataset final con 8019 registros y 3079 variables de las cuales 1 es la variable objetivo



softwareONE



Se utilizaron dos modelos uno de Máquinas de Soporte Vectorial (que es efectivo en espacios multidimensionales) y uno Bayesiano.

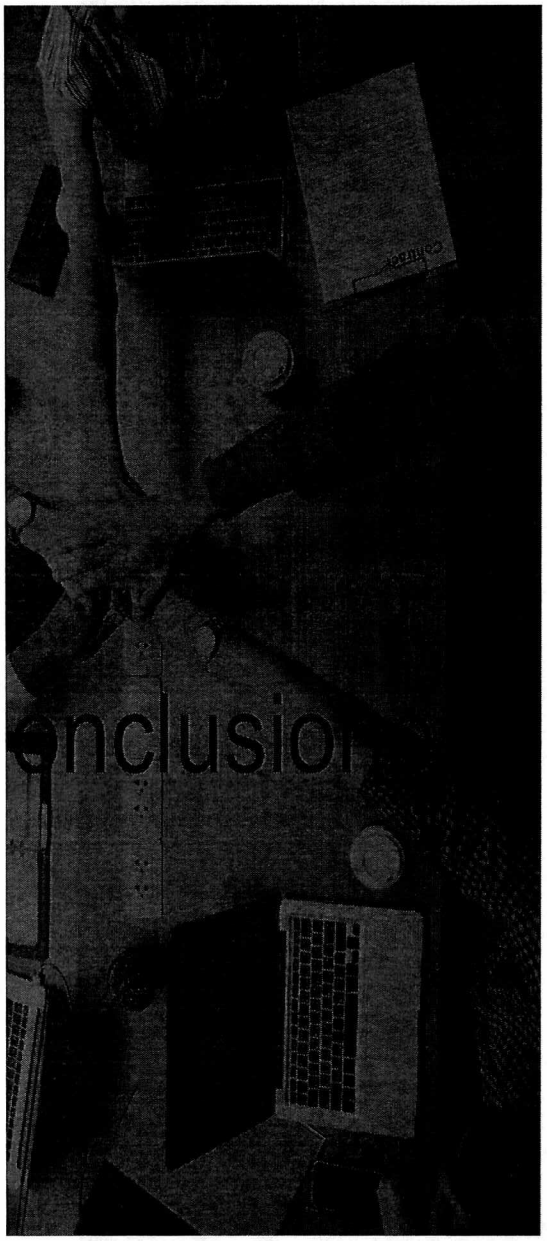
Maquinas de Soporte Vectorial

	precision	recall	f1-score	support
Estudiante en riesgo	0.87	0.75	0.81	257
Estudiante mayor de 18 años	0.96	1.00	0.98	52
Estudiante no trabajador	0.95	0.98	0.96	1242
Estudiante trabajador	0.84	0.70	0.76	53
micro avg	0.94	0.94	0.94	1604
macro avg	0.91	0.86	0.88	1604
weighted avg	0.93	0.94	0.93	1604

Bayesiano

	precision	recall	f1-score	support
Estudiante en riesgo	0.87	0.75	0.81	257
Estudiante mayor de 18 años	0.96	1.00	0.98	52
Estudiante no trabajador	0.95	0.98	0.96	1242
Estudiante trabajador	0.84	0.70	0.76	53
micro avg	0.94	0.94	0.94	1604
macro avg	0.91	0.86	0.88	1604
weighted avg	0.93	0.94	0.93	1604

softwareONE



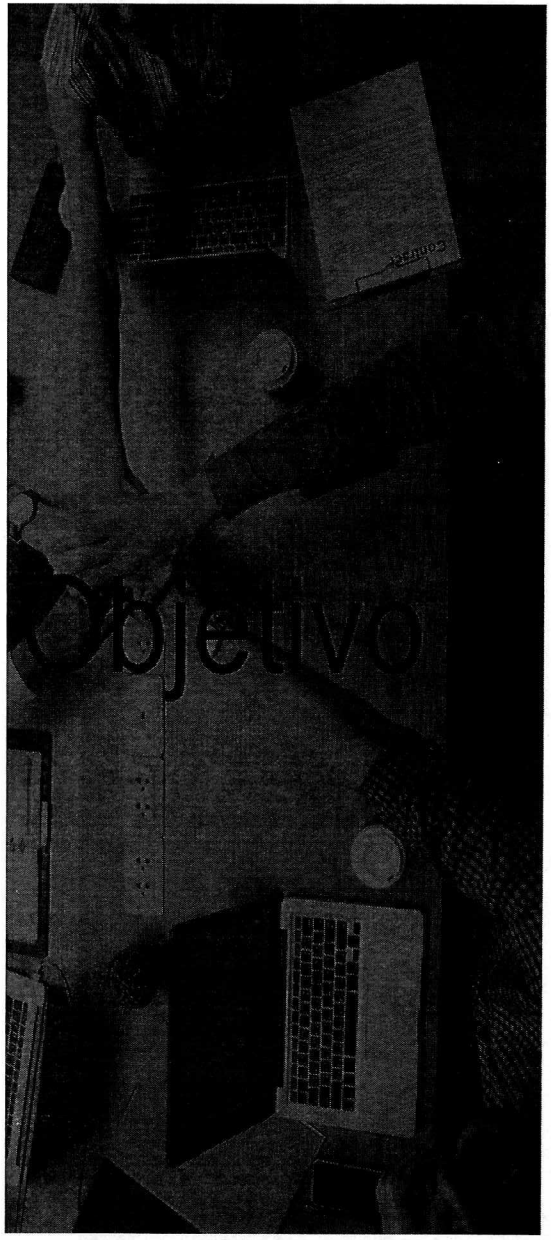
Como se puede observar en los reportes de clasificación y las matrices de confusión ambos modelos presentan los mismos resultados con unos problemas de precisión en la categoría de Estudiante en riesgo. El estudiante trabajador y el estudiante en riesgo tienen una precisión del 87% y 95% aproximadamente.

Anteriormente, se mencionaron inconsistencias en el dataset: los estudiantes masculinos en estado de embarazo, o los estudiantes categorizados como Estudiantes mayor de 18 años, cuando sus edades eran menores. En estos casos se podría tratar de obtener más datos, o revisar que la clasificación del conjunto de entrenamiento del dataset con la finalidad de obtener mejores resultados o tener una medición que cualifique el trabajo de los profesionales incluso cambiar la forma de la encuesta podría llegar a ser evaluado.

Es un modelo con un buen nivel de precisión, sin embargo, se sugiere tener la cualificación e incluso una mayor cantidad de data para mejorarlo. Puede ser puesto en producción con un tipo de aprendizaje on-line y monitorear con frecuencia su precisión.

softwareONE

thank
you
software 1



Crear un modelo que permita clasificar si un estudiante menor de edad está trabajando o está en riesgo laboral realizando un procesado de lenguaje natural sobre una pregunta abierta e incluyendo información socio-económica y académica. Se espera que con el perfeccionamiento del modelo – en el tiempo - la Secretaría de Educación logre una reducción de costos asociados a esta actividad.

softwareONE

almacenar, procesar, analizar y visualizar datos de cualquier tipo, volumen o velocidad, para que pueda descubrir nuevas oportunidades y adoptar medidas rápidamente.

Si quiere...	Utilice...
Almacenamiento de datos elástico totalmente administrado, con seguridad en todos los niveles de escalado y sin costo adicional.	SQL Data Warehouse
Plataforma de análisis totalmente administrada, rápida, sencilla, que permite la colaboración, basada en Apache® Spark™ y optimizada para Azure.	Azure Databricks
Un servicio de Hadoop y Spark en la nube totalmente administrado y respaldado por un acuerdo de nivel de servicio del 99,9 % para su empresa.	HDInsight
Un servicio de integración de datos para organizar y automatizar el traslado y la transformación de datos.	Data Factory
Desarrollo de inteligencia artificial abierto y elástico que abarca la nube y el perímetro.	Machine Learning
Procesamiento de flujo de datos en tiempo real procedente de millones de dispositivos IoT.	Azure Stream Analytics
Un servicio de análisis a petición totalmente administrado que se pague por trabajo, con seguridad, auditoría y soporte técnico de nivel empresarial.	Data Lake Analytics

218

Si quiere...	Utilice...
Motor de análisis de nivel empresarial como servicio.	Azure Analysis Services
Un servicio de uso de telemetría de hiperescala que recopila, transforma y almacena millones de eventos	Event Hubs
Servicio de exploración de datos muy escalable y rápido	Explorador de datos de Azure

4.2. Bases de Datos[7]

Servicios de bases de datos inteligentes totalmente administradas

Dondequiera que se encuentren sus datos, Microsoft Azure le ayudará a descubrir su potencial. Respalde un rápido crecimiento y ahorre más tiempo para innovación con una gama de servicios de base de datos totalmente administrada seguros y de nivel empresarial compatibles con motores de bases de datos de código abierto.

Los servicios de base de datos de Azure están totalmente administrados, lo que libera tiempo valioso para dedicarlo a la administración de la base de datos, a fin de poder centrarse en nuevas formas de sorprender a los clientes y de ofrecer oportunidades. El rendimiento empresarial con alta disponibilidad integrada significa que puede escalar rápidamente y abarcar una distribución global sin preocuparse por caros tiempos de inactividad. Además, los desarrolladores pueden beneficiarse de innovaciones líderes del sector, como la seguridad integrada con supervisión y detección de amenazas automáticas, ajuste automático para un rendimiento mejorado y distribución global llave en mano. Lo más

[7] https://azure.microsoft.com/es-es/product-categories/databases/

importante de todo es que la inversión está protegida por un SLA de respaldo financiero.

Independientemente de lo que cree, le ayudaremos a comercializarlo rápidamente, distribuirlo a nivel general y administrarlo con facilidad y confianza.

Si quiere...	Utilice...
Base de datos multimodelo distribuida globalmente que admite opciones NoSQL y ofrece un rendimiento y contratos de nivel de servicio excepcionales y Acuerdos de Nivel de Servicio excepcionales	Azure Cosmos DB
Una base de datos relacional totalmente administrada que se aprovisione con rapidez, admita escalado sobre la marcha e incluya inteligencia y seguridad integradas	Azure SQL Database
Una base de datos relacional MySQL escalable y totalmente administrada, con alta disponibilidad y seguridad integrada sin costo adicional	Azure Database for MySQL
Una base de datos relacional PostgreSQL escalable y totalmente administrada , con alta disponibilidad y seguridad integrada sin costo adicional	Azure Database for PostgreSQL
Hospedaje de aplicaciones empresariales de SQL Server en la nube	SQL Server en Virtual Machines

220

Si quiere...	Utilice...
Almacenamiento de datos elástico totalmente administrado, con seguridad en todos los niveles de escalado y sin costo adicional	SQL Data Warehouse
Ayuda para migrar sus bases de datos a la nube sin cambios de código en las aplicaciones	Azure Database Migration Service
Acceso a datos consistente, de baja latencia y alto rendimiento para implementar aplicaciones rápidas y escalables	Azure Cache for Redis
Un almacén claves-valores NoSQL para desarrollo rápido que usa conjuntos de datos semiestructurados masivos	Table Storage
Una base de datos relacional MariaDB escalable y totalmente administrada, con alta disponibilidad y seguridad integrada sin costo adicional	Azure Database for MariaDB

4.3. Data Factory[8]

Capacidades del Servicio:

Obtenga acceso e ingiera datos con los conectores integrados

Migre datos de orígenes locales y en la nube a un almacén de datos centralizado en la nube para su posterior análisis con la actividad Copiar datos en una canalización de datos.

[8] https://docs.microsoft.com/es-es/azure/data-factory/

221

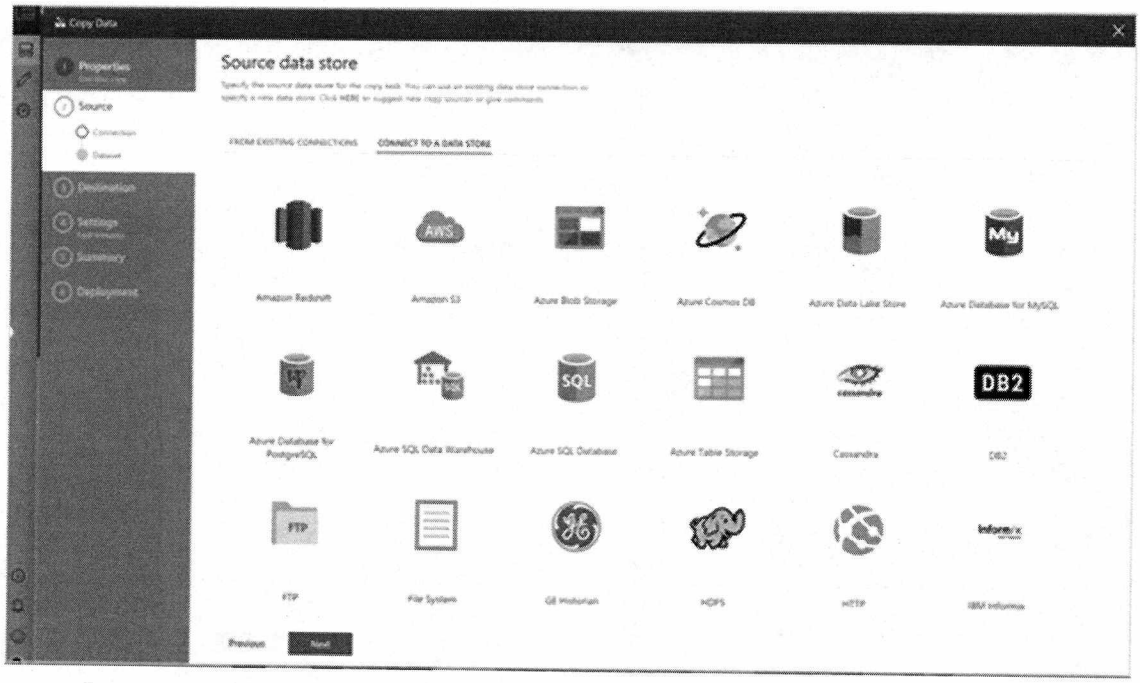


Imagen conectores

Compile flujos de datos escalables con la interfaz de usuario sin código o escriba su propio código

Cree integración de datos y transforme e integre fácilmente el aprendizaje automático y el procesamiento de macro datos con la interfaz visual.

222

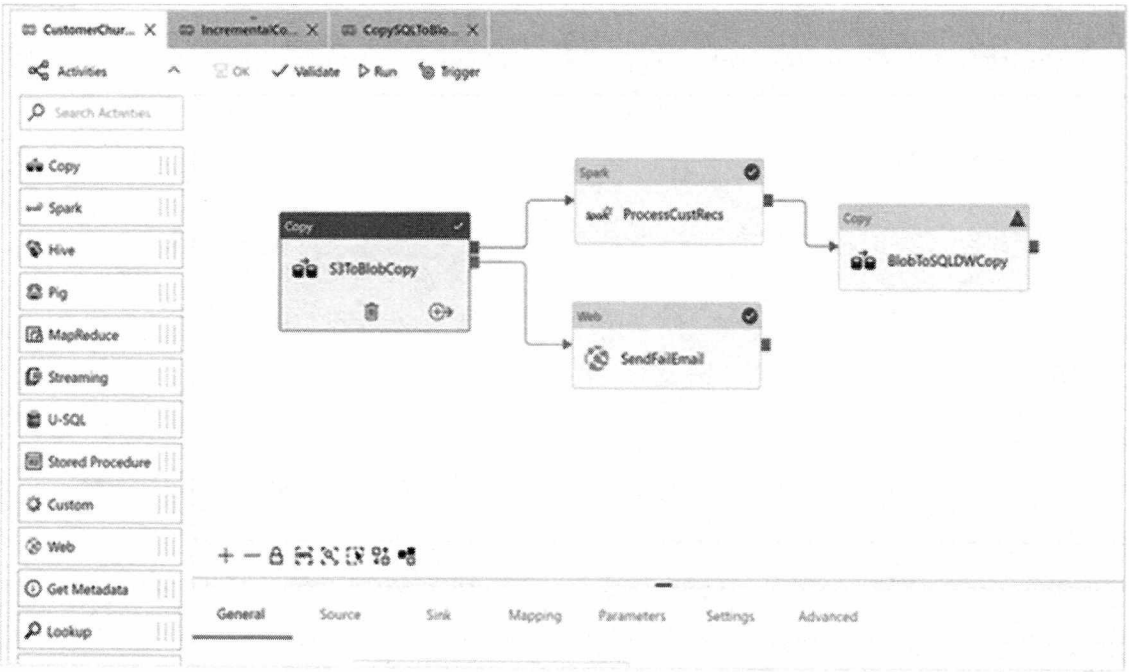


Imagen flujo de datos

Programe, ejecute y supervise sus canalizaciones

Invoque canalizaciones con programación basada en desencadenadores y a petición. Supervise visualmente la actividad de las canalizaciones con el historial de canalizaciones y registros, y realice el seguimiento de los orígenes de errores.

223

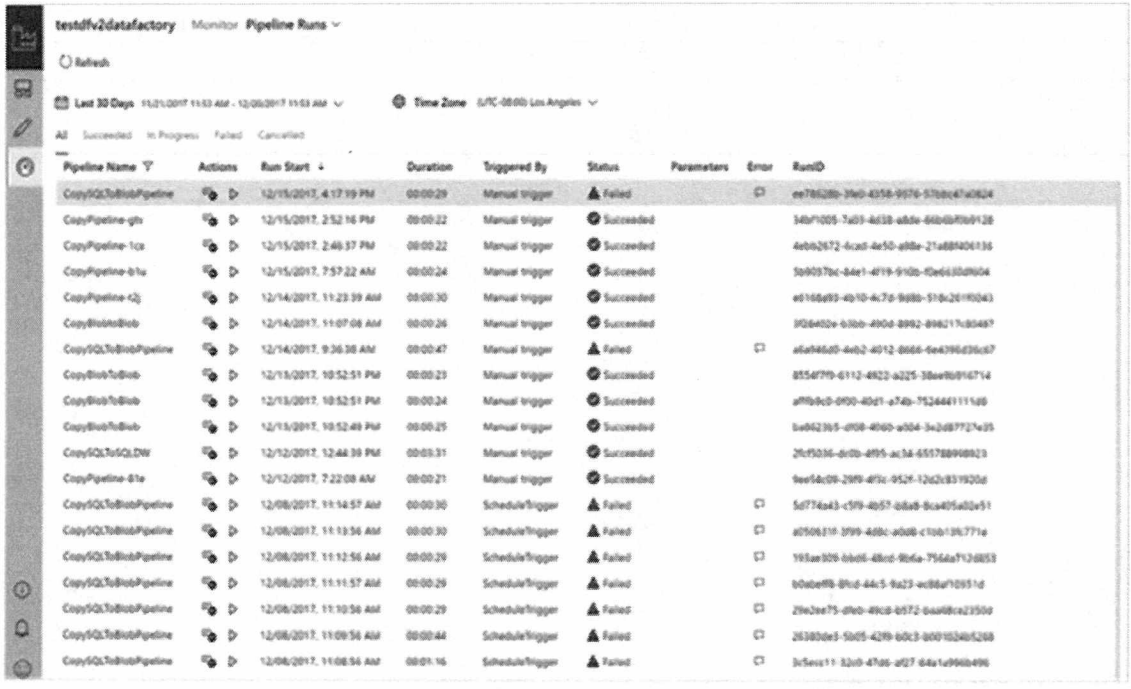
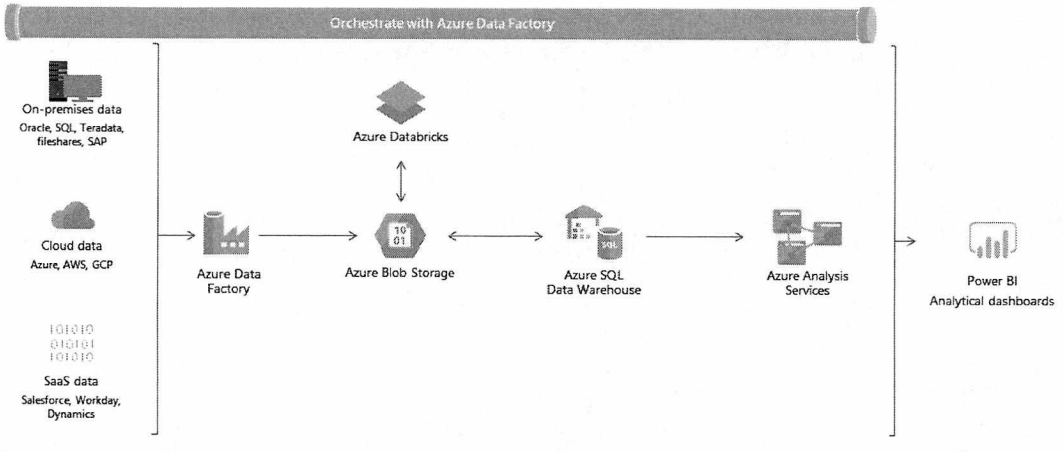


Imagen ejecución canalizaciones (pipelines)

Integre en un solo almacén de datos todas sus fuentes de datos en sitio, aplicaciones SAAS o data en archivos semiestructurados.



Azure Data Factory also integrates with other services such as Azure HDInsight, Azure Data Lake, and Azure Machine Learning to allow customers to tailor the above architecture to meet their unique needs.

224

4.4. DataBase[9]

Migre bases de datos en masa y escale hasta 100 TB

Consiga almacenamiento rápido y dinámico con SQL Database a hiperescala, la nueva tecnología de escalado de bases de datos de alto rendimiento. El nivel de servicio Hiperescala permite que SQL Database se adapte a sus cargas de trabajo y optimice los recursos a petición.

Aprovisione uno o varios nodos de proceso más que pueden atender la carga de trabajo de solo lectura y úselos como espera activa en caso de conmutación por error.

Lleve a cabo las operaciones en un tiempo constante, al margen del tamaño de la operación de datos.

Escale los recursos de proceso y almacenamiento por separado y en nada de tiempo, sin sacrificar el rendimiento.

Desarrolle aplicaciones seguras en la nube

Utilice protección inteligente y características extraordinarias de seguridad y privacidad para:

Controlar el acceso a su base de datos con autenticación multifactor.

Dejar los datos confidenciales cifrados mientras se utilizan con Always Encrypted.

Supervisar la base de datos con Advanced Threat Protection para evitar posibles amenazas y vulnerabilidades.

SQL Database cumple los estándares más exigentes, como GDPR, ISO/IEC 27001/27002, FedRAMP/FISMA, SOC, HIPAA y PCI DSS.

225

[9] https://docs.microsoft.com/es-es/azure/sql-database/

Asegure la alta disponibilidad con tres réplicas de acceso frecuente y conmutación por error automática integrada con un acuerdo de nivel de servicio que garantiza una disponibilidad del 99,99 %.

Optimice el rendimiento con inteligencia integrada

Mejore el uso de memoria, el procesamiento, la latencia y el rendimiento de las consultas.

Ajuste inteligente del rendimiento utiliza inteligencia integrada con aprendizaje automático y tecnologías adaptables para optimizar el rendimiento de las bases de datos de forma constante en tiempo real. Obtenga recomendaciones de ajuste, conclusiones y mejoras de rendimiento de forma automática.

Procesamiento de consultas inteligente mejora el rendimiento de las cargas de trabajo con un esfuerzo de implementación mínimo. Incluye características como el procesamiento adaptable de consultas, que adapta las estrategias de optimización a las condiciones en tiempo de ejecución de la carga de trabajo de la aplicación, y procesamiento aproximado de consultas, diseñado para agregar conjuntos de datos de gran tamaño cuando la capacidad de respuesta es crucial.

4.5. Vista de Datos

Son tablas maestras y especializadas en el almacenamiento de información de un área o de un tema específico, son alimentadas desde las fuentes de datos y reguladas por las reglas de negocio que brindaran información detallada, escalable, que se actualiza de forma constante y se pobla resultado del proceso de ingesta de información.

4.6. PowerBI

Power BI es una solución de análisis empresarial que permite visualizar los datos y compartir información con toda la organización, o insertarla en su aplicación o sitio web. Conéctese a cientos de orígenes de datos y dé vida a sus datos con los paneles e informes dinámicos.[10]

5. Analisis de Arquitectura Inicial

5.1. Información

La arquitectura inicial está basada en Dynamics 365 más un integrador desarrollado por un ISV que integra la información de las fuentes de datos concentradas en archivos de Excel para la carga en la base de datos de la aplicación.

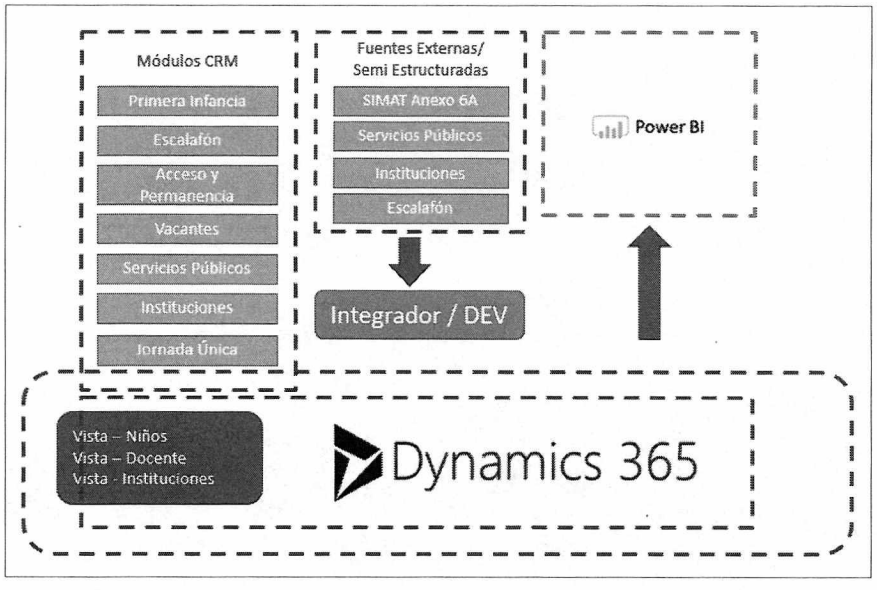


Diagrama de arquitectura inicial

[10] https://powerbi.microsoft.com/es-es/what-is-power-bi/

227

Diagnóstico de la arquitectura:

En levantamiento de información con la Secretaría de Educación se evidencia la necesidad de tener un sistema que le de flexibilidad, agilidad y seguridad a la inclusión de nuevas fuentes de información.

Adicionalmente se evidencia que el uso de la base de datos de Dynamics va en ascenso dado que la información alojada allí corresponde a los cargues de SIMAT que se realizaran una vez al mes para el consumo de la información por parte del área de jornada única y primera infancia.

El costo relacionado con el crecimiento de la base de datos en Dynamics se proyectaba con incremento sostenido por el número de años de histórico que la secretaria desea tener como base del análisis. Haciendo que la solución se torne onerosa frente a soluciones de almacenamiento en base de datos y con mayor flexibilidad en la ingesta de los mismos.

Precios de referencia:

https://dynamics.microsoft.com/en-us/pricing/

https://azure.microsoft.com/es-es/pricing/details/sql-database/single/

5.2. Sistemas de información

Los sistemas de información se definen como un conjunto de elementos orientados al tratamiento y administración de datos e información, organizados y listos para su uso posterior, generados para cubrir una necesidad u objetivo[11]

- Dynamics 365
- Humano

228

[11] https://es.wikipedia.org/wiki/Sistema_de_informaci%C3%B3n

- SIMAT - Sistema Integrado de Matrícula:
- Integrador

5.3. Administración de datos

Basado en los roles de seguridad de Dynamics 365, en la presentación de datos relacionados con las áreas usuarias y su actualización mediante procesos manuales y automatizados para la ingesta fija de datos y visualización de informes que les permitan tener información de valor para la toma de decisiones.

5.4. Estructura de datos SIMAT

Este diagrama corresponde al Modelo entidad relación, de las entidades utilizadas dentro de SIMAT, que se utilizó en Dynamics 365/CRM.

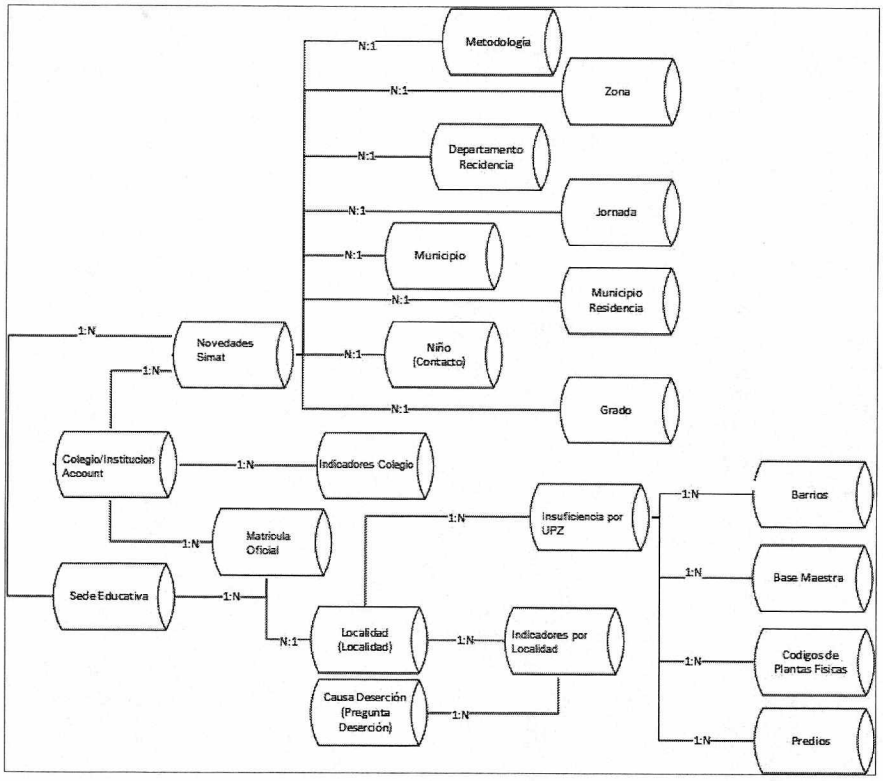


Diagrama 1. SIMAT - Dynamics 365.

229

5.5. Estructura de datos Acceso y permanencia

Los siguientes diagramas corresponden al Modelo entidad relación, de las entidades utilizadas dentro de las aplicaciones de Dynamics 365/CRM. Para el modulo de acceso y permanencia (bienestar alimentación, bienestar movilidad, base maestra – proyectos y mejoramientos, dotaciones).

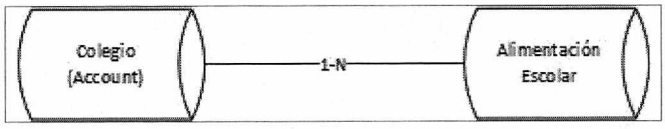


Diagrama 2. Bienestar Alimentación - Dynamics 365.

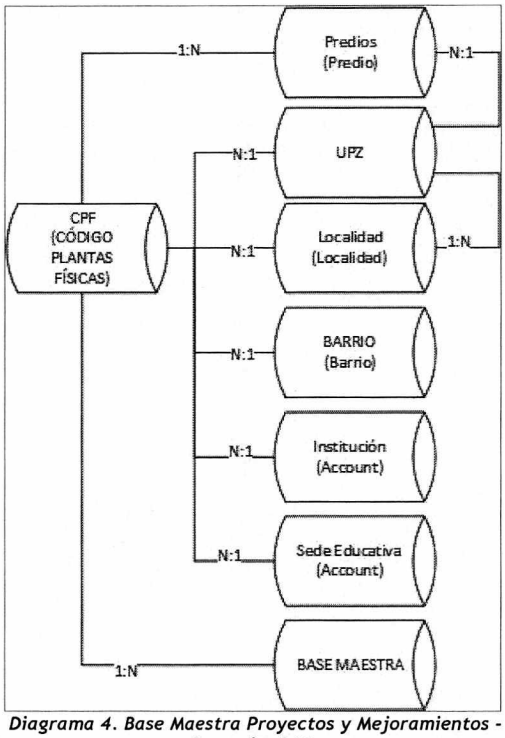


Diagrama 4. Base Maestra Proyectos y Mejoramientos - Dynamics 365.

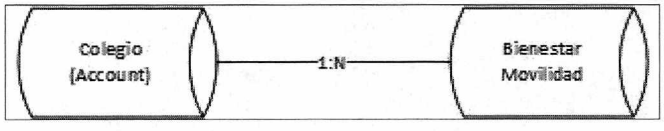


Diagrama 3. Bienestar Movilidad - Dynamics 365.

230

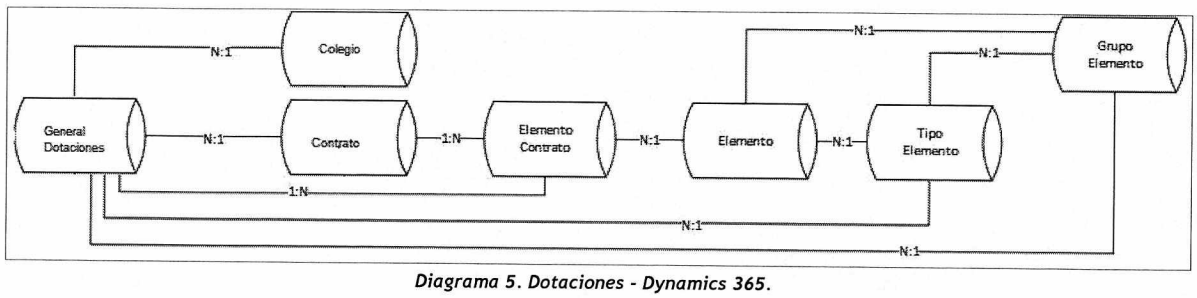


Diagrama 5. Dotaciones - Dynamics 365.

5.6. Estructura de datos Escalafón

Este diagrama corresponde al Modelo entidad relación, de las entidades utilizadas para la aplicación de Escalafón, que se utilizó en Dynamics 365/CRM.

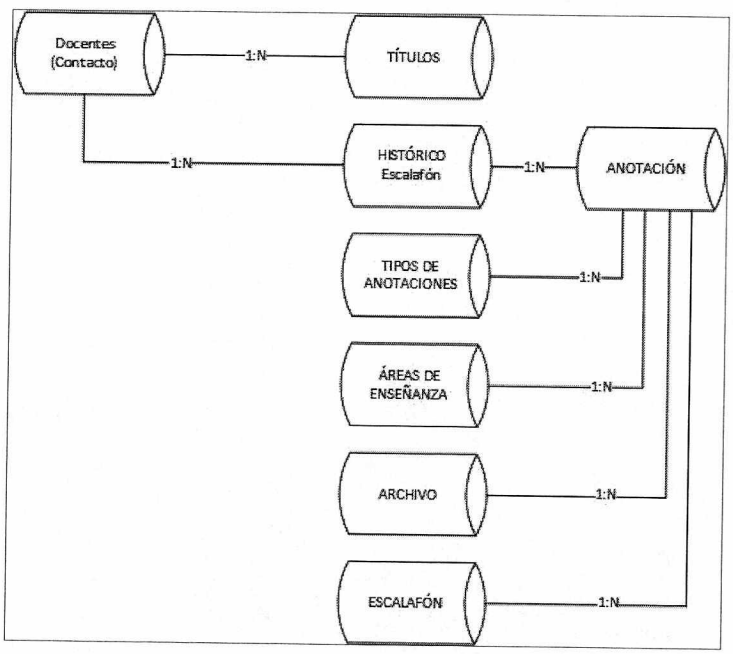


Diagrama 6. Escalafón - Dynamics 365

231

5.7. Estructura de datos Servicios Públicos

Este diagrama corresponde al Modelo entidad relación, de las entidades utilizadas para la aplicación de Servicios Públicos, que se utilizó en Dynamics 365/CRM.

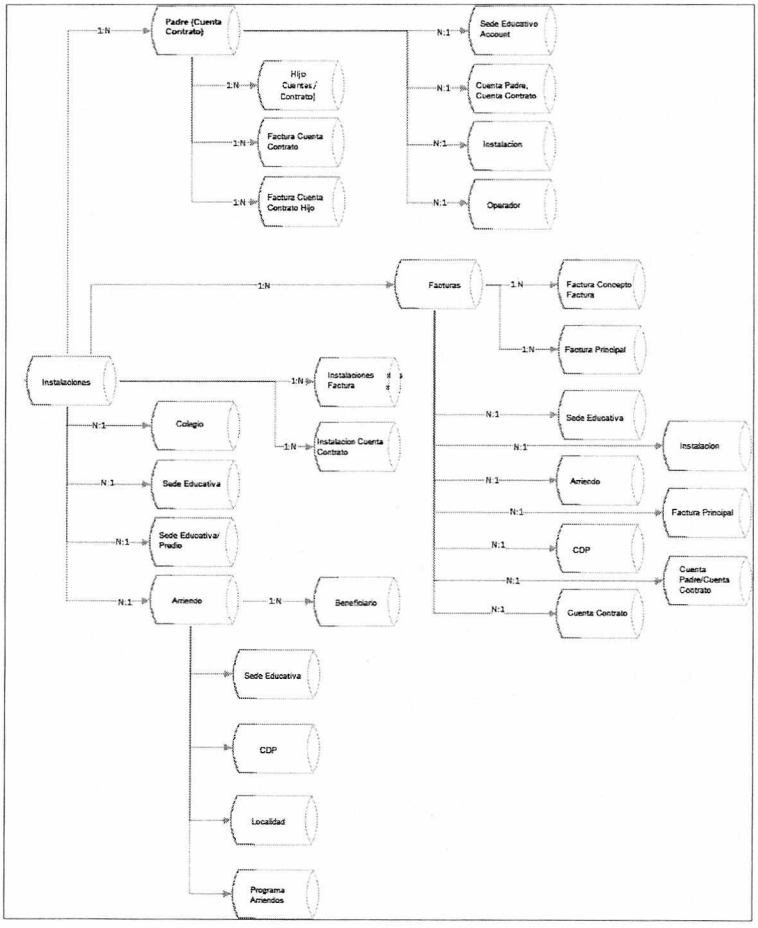


Diagrama 7. Servicios Públicos - Dynamics 365.

232

5.8. Estructura de datos Instituciones

Este diagrama corresponde al Modelo entidad relación, de las entidades utilizadas para la aplicación de Instituciones, que se utilizó en Dynamics 365/CRM.

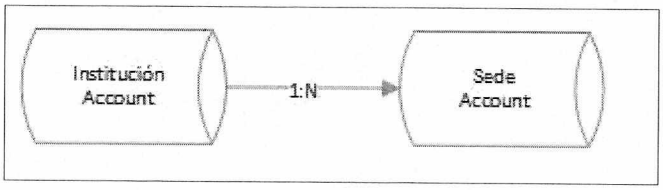


Diagrama 8. Institución - Dynamics 365.

6. Diseño Arquitectura Nueva

6.1. Arquitectura de la solución

A continuación, presentamos el diseño conceptual y el diseño definitivo de la solución que comprende la estructuración y origen de los datos para el sistema de información que la Secretaría usara como almacén de datos.

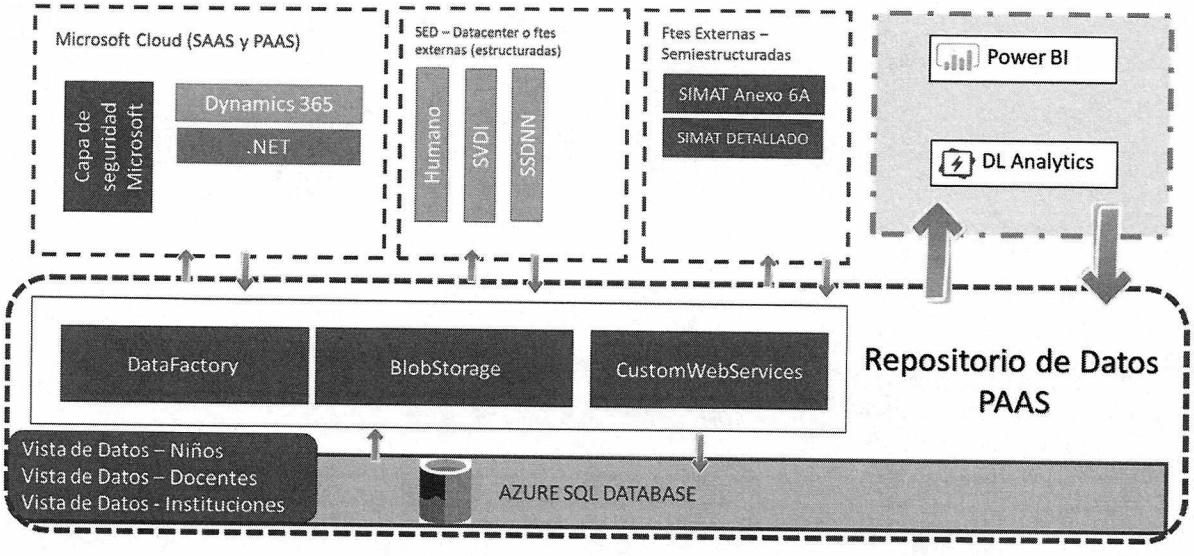


Diagrama 9. Solución Conceptual.

233

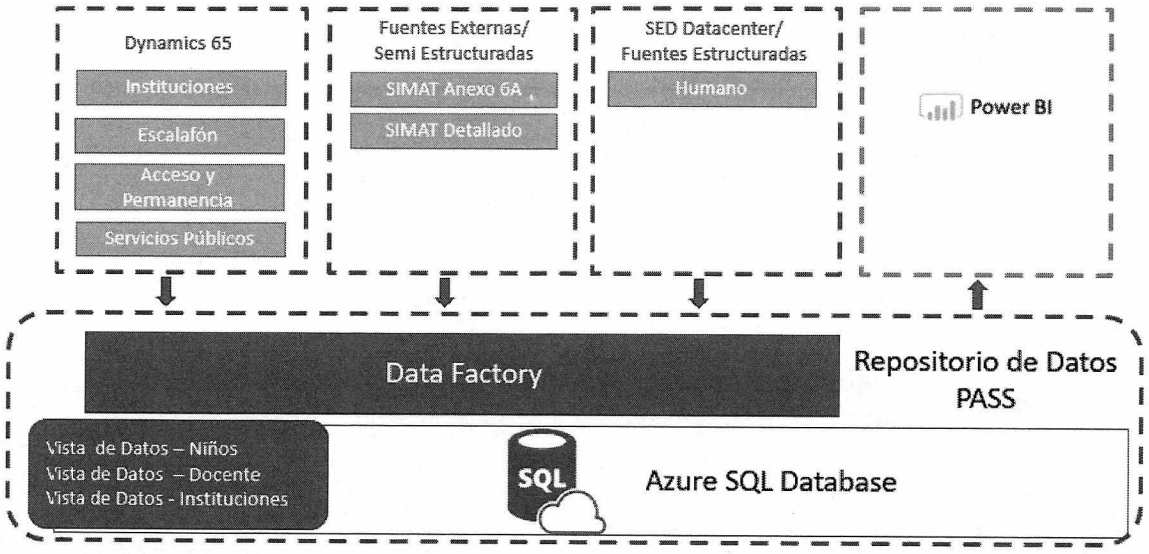


Diagrama 10. Solución Actual.

La solución propuesta comprende el uso del servicio de análisis de Microsoft Azure como base de crecimiento para la inclusión de 3 fuentes en este contrato y abriendo la posibilidad a futuro de la inclusión de nuevas fuentes de información que complementaran las vistas de datos diseñados.

En primera instancia Data Factory como servicio de integración (ETL – Extracción, Transformación y Carga) de información con conectores prefabricados por Microsoft o procesos almacenados que podrán personalizar los procesos de transformación de la información conforme lo necesite la fuente.

En segunda instancia un servicio de base de datos basado en SQL Server en la modalidad de plataforma como servicio en donde la Secretaría tendrá la posibilidad de escalar sus capacidades en cuanto al procesamiento y almacenamiento según su necesidad y disminuir el impacto de administración sobre los elementos de infraestructura típicos de una solución en sitio, ganando tiempo y optimizando su operación al tratamiento y mejoramiento de los datos en relación con los procesos de mejora continua que tiene la entidad.

En tercera instancia la visualización de informes a través de PowerBI para dar uso de la información a los usuarios funcionales conservando las características de seguridad y beneficios técnicos propias del modelado de informes con esta herramienta tecnológica.

Como parte del crecimiento de la plataforma se plantea la posibilidad conceptual de poder integrar más aplicaciones a través de web services personalizados y alojados en la plataforma de aplicaciones de Azure y adicionalmente usar el motor de *"análisis services"* para las necesidades crecientes de la entidad.

6.2. Componentes

Esta solución comprende los siguientes componentes:

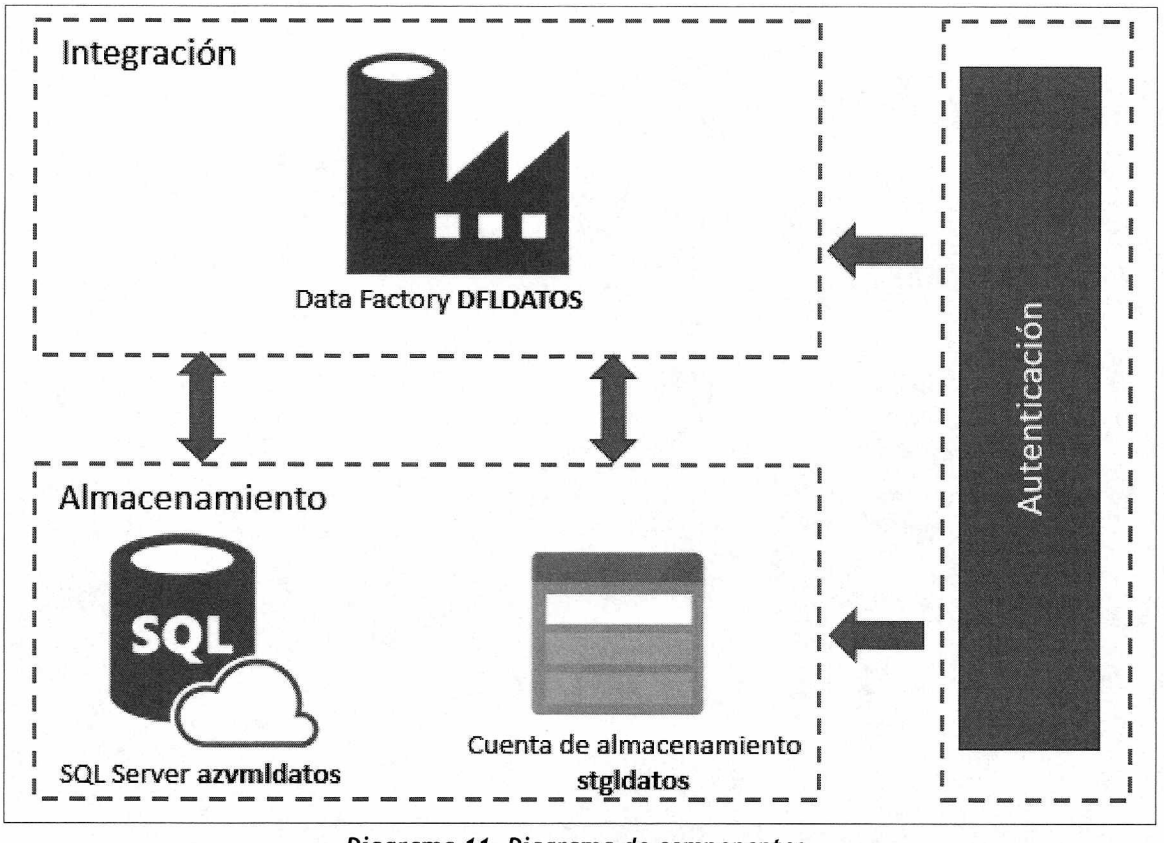
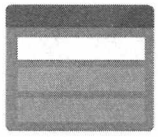


Diagrama 11. Diagrama de componentes.

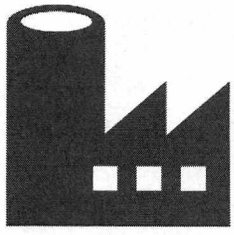
- **Storage Account:** Una cuenta de Azure Storage contiene todos los objetos de datos de almacenamiento de Azure. Este componente se encuentra en el grupo de recursos RG-EUS2-VM-LDATOS, con nombre stgldatos, versión (Storage uso general v1).



- **Servicio SQL Server:** SQL Database es un servicio administrado de base de datos relacional de uso general de Microsoft Azure que admite estructuras como datos relacionales, JSON, espacial y XML. SQL Database ofrece un rendimiento escalable de modo dinámico. Este componente se encuentra en el grupo de recursos RG-EUS2-VM-LDATOS, con nombre azvmldatos, base de datos AZDBLDATOS (ubicación EE.UU.2, Versión Microsoft SQL Azure (RTM) - 12.0.2000.8 Dec 19 2018 Copyright (C) 2018 Microsoft Corporation).



- **Data Factory:** Azure Data Factory es un servicio de integración de datos. Que permite extraer y transformar datos a escala de nube. Se encuentra en el grupo de recursos RG-EUS2-VM-LDATOS, ubicación Este de EE.UU.2, Data Factory v2, con nombre DFLDATOS.



236

La comunicación entre la cuenta de almacenamiento y Data Factory es bidireccional, porque se integran fuentes almacenadas allí y el log de registros no cargados en el proceso[12].

Entre SQL Server y Data Factory la comunicación es bidireccional debido que se integran los proceso sincronización de datos y la ejecución de procedimientos almacenados.

Todas las transferencias de datos entre los servicios de movimiento de datos de Data Factory y un almacén de datos en la nube se realizan a través del canal seguro HTTPS o TLS.

Microsoft. (2018). Consideraciones de seguridad para el movimiento de datos en Azure Data Factory. 2019, de Microsoft Sitio web: https://docs.microsoft.com/es-es/azure/data-factory/data-movement-security-considerations

6.3. Ventajas para el negocio

- Tecnología con conocimiento publicado a través de los portales de documentación del fabricante.
- Soporte directo con el fabricante.
- Disponibilidad de la plataforma según las políticas de Azure 99.9%
- Disminuir el impacto de administración de las soluciones de datos.
- Posibilidad de inclusión de cualquier fuente nueva de datos estructurada, semiestructurada o no estructurada en el futuro.

6.4. Sistemas de información

Dynamics 365:

Sistema de información que contiene la digitalización de los procesos de la subsecretaria de accesos y permanencia, la dirección de servicios administrativos, escalafón y la oficina de personal

237

[12] La descripción detallada de los procesos se encuentra en los manuales técnicos (Manual Técnico Simat, Manual Técnico XRM, Manual Técnico DUEB, Manual Técnico Humano).

Humano:

Sistema de información que contiene el escalafón de docentes públicos y las novedades de ausencias en la planta de educación del distrito.[13]

SIMAT - Sistema Integrado de Matrícula:

El sistema integrado de matrícula SIMAT es una herramienta que permite organizar y controlar el proceso de matrícula en todas sus etapas, así como tener una fuente de información confiable y disponible para la toma de decisiones.

Es un sistema de gestión de la matrícula de los estudiantes de instituciones oficiales que facilita la inscripción de alumnos nuevos, el registro y la actualización de los datos existentes del estudiante, la consulta del alumno por Institución y el traslado a otra Institución, entre otros.[14]

6.5. Por qué un Repositorio de datos, características evaluadas:

Datos: Un data warehouse, en su base de datos, sólo almacena datos que han sido modelados o estructurados; mientras que la base de datos del repositorio de datos no hace excepción de datos. Lo almacena todo, estructurado, semiestructurado y no estructurado.

Procesamiento: Antes de que una empresa pueda cargar datos en un data warehouse, primero debe darles forma y estructura, es decir, los datos deben ser modelados. Eso se llama schema-on-write. Con un repositorio de datos, sólo se cargan los datos sin procesar, tal y como están, y cuando esté listo para usar los datos, es cuando se le da forma y estructura. Eso se llama schema-on-read. Dos enfoques muy diferentes, lo cual permite al repositorio ser más flexible.

[13] Definición del sistema no disponible.
[14] https://www.mineducacion.gov.co/1621/article-168883.html

Almacenamiento: Una de las principales características de las tecnologías de big data, es que el costo de almacenamiento de datos es relativamente bajo en comparación con el de un data warehouse. La principal razón es: que el pago que en Azure para el repositorio de datos se realiza es por el servicio; mientras que para el data warehouse, existen una serie de planes acorde a la cantidad de almacenamiento (Teras) que se vayan a usar, lo cual hace que este segundo sea más costoso por la limitante del almacenamiento atado a un costo fijo mensual, el data warehouse su usa para soluciones muy robustas en cuanto al almacenamiento se refiere; igualmente el data warehouse se instala en máquinas robustas que puedan soportar esta cantidad en Teras de almacenamiento, lo que también eleva su costo y mantenimiento.

Agilidad: Un data warehouse, es un repositorio altamente estructurado, por definición. No es técnicamente difícil cambiar la estructura, pero puede tomar mucho tiempo dado por todos los procesos de negocio que están vinculados a este. Un repositorio de datos como el de Azure, por otro lado, carece de la estructura de un data warehouse, lo que da la posibilidad a los desarrolladores y a los científicos de datos la capacidad de configurar y reconfigurar fácilmente y en tiempo real sus modelos, consultas y aplicaciones.

6.6. Administración de Datos

Datos operacionales: roles de seguridad administrados por las aplicaciones como Dynamics 365 y las aplicaciones web desarrolladas.

Datos de las tablas en la base de datos de administrarán con la seguridad y manejo de usuarios de SQL, más Power BI para compartir los reportes entre los usuarios.

239

6.7. Vistas de datos

De acuerdo con la estructura de las fuentes de datos, la información se agrupa en tres vistas de datos: Instituciones, Estudiantes y Docentes.

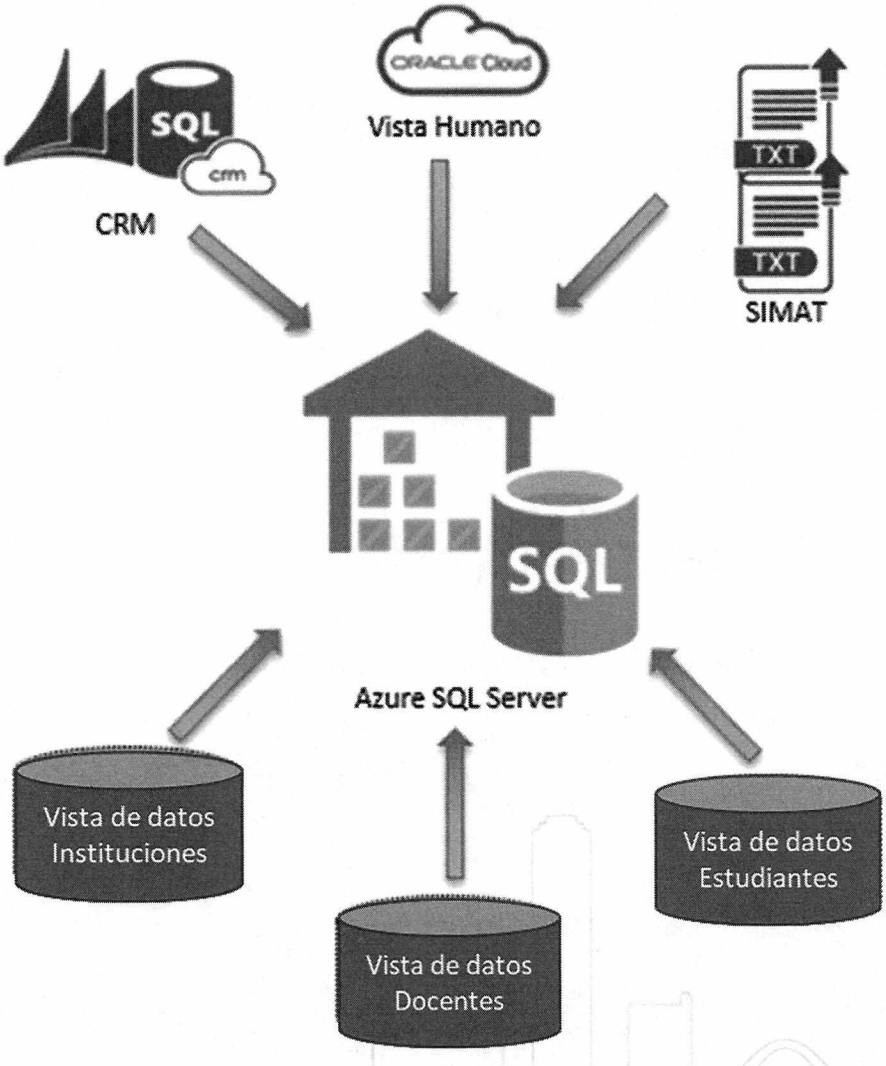


Diagrama 12. Vista de datos Repositorio SED

240

6.7.1. Vista de Datos Docentes

Contiene información relacionada con el escalafón de los docentes.

Las fuentes son:

- CRM para sector privado.
- Humano para sector público.

De acuerdo con el modelo de la vista de datos, se puede establecer que el repositorio de datos contendrá la siguiente información del docente:

- Datos personales generales del docente (Nombre, Número de documento, tipo de documento, dirección de residencia, correo electrónico, número de teléfono)-(CRM-Humano).
- Vinculaciones y retiros del docente (Humano).
- Solicitudes de escalafón del docente (CRM).
- Institución actual en la que se desempeña como docente (Humano).
- Sede de la institución actual en la que se desempeña el docente (Humano).
- Jornada actual (Humano).

241

6.7.4. Relación de vista de datos en el Repositorio Datos.

Con la información que se obtiene de fuentes de datos que se integran al repositorio de datos de SED, se define que la relación entre las tres vistas de datos es la siguiente.

Instituciones y docentes se relacionan por el dane12 de la institución y dane12 de la sede.

Instituciones y estudiantes se relacionan por el dane12 de la institución, Dane12 de la sede.

La relación entre docente y estudiantes no tiene atributos en común.

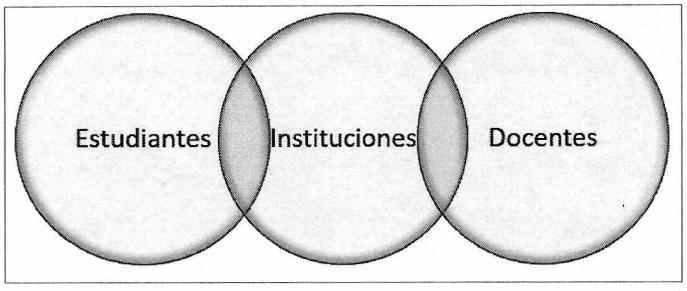


Diagrama 18. Relación vista de datos

6.7.5. Ejecución de procesos por Vista de datos.

Para el adecuado funcionamiento del proceso de integración y teniendo en cuenta las reglas para cada fuente, las primeras actividades que se deben ejecutar, son las asociadas a la vista de datos de instituciones y posteriormente las asociadas a estudiantes o docentes.

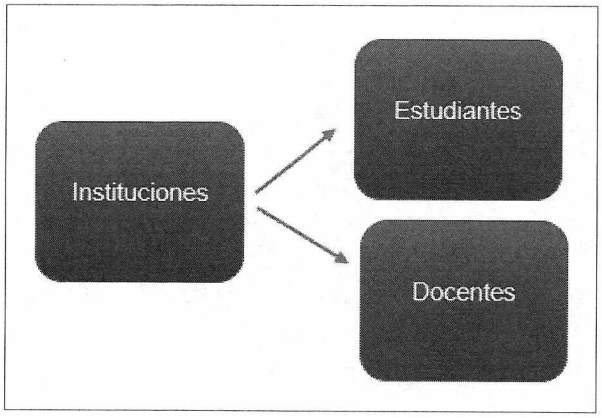


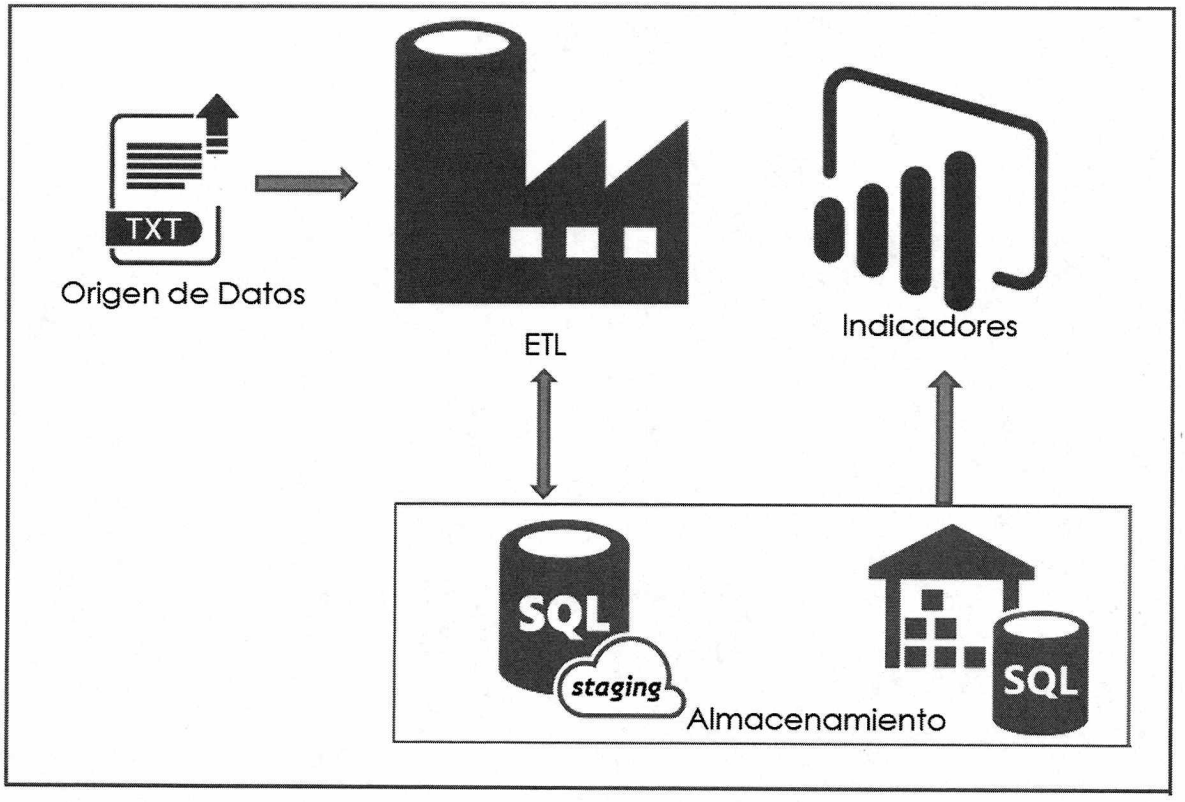
Diagrama 19. Ejecución de Procesos por vista de datos

Nota: el detalle de los procesos que se deben ejecutar para la integración están definidos en el documento técnico simat, documento técnico xrm y documento técnico humano.

7. Desarrollo del proyecto

7.1. Integración Fuente Simat

250



15

7.1.1. Definición de la fuente:

Simat: Simat es un sistema de información del Ministerio de Educación Nacional MEN. Al no contar con un proceso que permita la conexión directa (web services), se genera un archivo en formato txt., que contiene la información de la matrícula de los estudiantes del distrito y se genera con una periodicidad mínima de una semana.

Nota: El procedimiento de carga del archivo de SIMAT al blob storge de Azure está descrito en el Documento Técnico de SIMAT.

7.1.2. Procesos:

[15] Los indicadores se describen de forma general en el numeral 7.6.

Para la integración de la fuente se desarrolla proceso de Extracción, Transformación y Carga.

7.1.2.1. Extracción

Para realizar la extracción de los datos se configura el servicio de Azure Data Factory. Para que se mapee la información del archivo de texto Simat.txt, se estructuren los campos que lo componen y guarde los datos en las columnas definidas dentro de la tabla de staging, conservando la misma estructura de la fuente.

Tabla 1: Estructura del archivo SIMAT

ATRIBUTO	TIPO DATO	ATRIBUTO	TIPO DATO
ANO_INF	varchar(50)	ETNIA	varchar(50)
MUN_CODIGO	varchar(50)	RES	varchar(50)
CODIGO_DANE	varchar(50)	INS_FAMILIAR	varchar(255)
DANE_ANTERIOR	varchar(50)	TIPO_JORNADA	varchar(50)
CONS_SEDE	varchar(50)	CARACTER	varchar(50)
TIPO_DOCUMENTO	varchar(50)	ESPECIALIDAD	varchar(50)
NRO_DOCUMENTO	varchar(50)	GRADO	varchar(50)
EXP_DEPTO	varchar(50)	GRUPO	varchar(50)
EXP_MUN	varchar(50)	METODOLOGIA	varchar(50)
APELLIDO1	varchar(50)	MATRICULA_CONTRATADA	varchar(50)
APELLIDO2	varchar(50)	REPITENTE	varchar(50)
NOMBRE1	varchar(50)	NUEVO	varchar(50)
NOMBRE2	varchar(50)	SIT_ACAD_ANO_ANT	varchar(50)
DIRECCION_RESIDENCIA	varchar(255)	CON_ALUM_ANO_ANT	varchar(50)
TEL	varchar(50)	FUE_RECU	varchar(50)
RES_DEPTO	varchar(50)	ZON_ALU	varchar(50)
RES_MUN	varchar(50)	CAB_FAMILIA	varchar(50)
ESTRATO	varchar(50)	BEN_MAD_FLIA	varchar(50)
SISBEN	varchar(50)	BEN_VET_FP	varchar(50)
FECHA_NACIMIENTO	varchar(50)	BEN_HER_NAC	varchar(50)
NAC_DEPTO	varchar(50)	CODIGO_INTERNADO	varchar(50)
NAC_MUN	varchar(50)	CODIGO_VALORACION_1	varchar(50)
GENERO	varchar(50)	CODIGO_VALORACION_2	varchar(50)
POB_VICT_CONF	varchar(50)	NUM_CONVENIO	varchar(50)
DPTO_EXP	varchar(50)	PER_ID	varchar(50)
MUN_EXP	varchar(50)	APOYO_ACADEMICO_ESPECIAL	varchar(50)
PROVIENE_SECTOR_PRIV	varchar(50)	SRPA	varchar(50)
PROVIENE_OTR_MUN	varchar(50)	FECHA_INSERCION	datetime
TIPO_DISCAPACIDAD	varchar(50)	FECHA_CORTE	datetime
CAP_EXC	varchar(50)	ID_CARGA	uniqueidentifier

7.1.2.2. Transformación:

252

En este proceso se aplican las reglas de negocio requeridas para la integración de SIMAT con el repositorio de datos.

El registro de un estudiante se considera único por número de documento.

El estudiante debe pertenecer a un colegio y/o una sede.

El registro del estudiante debe cumplir con criterios de completitud asociados a los datos maestros (Etnia, discapacidad, población vulnerable, jornada, grado, metodología, carácter, especialidad, tipo de documento etc.).

Los registros que no se cargan a la tabla destino de los estudiantes en el repositorio de datos, por no cumplir con cualquiera de las reglas anteriormente descritas, se almacenarán en tablas de staging, con la descripción del motivo de rechazo[16].

7.1.2.3. Carga:

En este proceso se tienen en cuenta únicamente los registros que cumplen con las reglas de negocio. Cuando se presentan cambios en la información del estudiante se actualizarán, si no existe será insertado en la tabla destino.

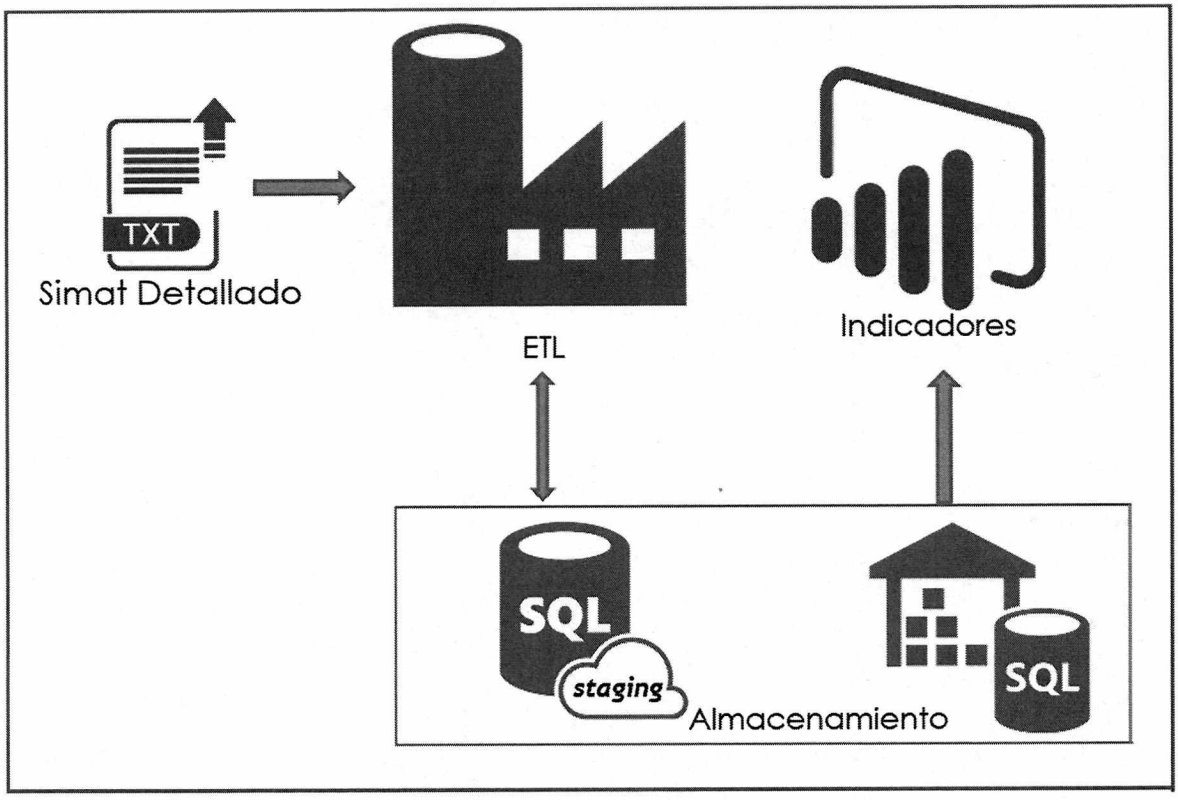
7.1.2.4. Programación de ejecución de procesos

Para la integración de este proceso se requiere la creación de un flujo de trabajo dentro del servicio de Azure Data Factory, la programación de ejecución del proceso depende de la periodicidad de generación del

[16] Los nombres de los procesos, las tablas de origen destino y rechazo, están especificados en el Documento técnico de SIMAT y en el Diccionario de datos.

archivo simat.txt y la previa ejecución de los procesos de integración de la vista de datos instituciones[17].

7.2. [18]Integración Fuente Simat Detallado.



7.2.1. Definición de la fuente.

Simat detallado: Simat es un sistema de información del Ministerio de Educación Nacional MEN. Al no contar con un proceso que permita la conexión directa (web services), se genera un archivo en formato txt., que contiene la información de las novedades de los estudiantes matriculados

[17] La programación de la ejecución o la ejecución manual de los procesos está documentada en el Manual Técnico Simat.
[18] Los indicadores se describen de forma general en el numeral 7.6.

en el distrito, (cambios de colegio, sede, jornada, grado etc.); que se genera diariamente.

Nota: El procedimiento de carga del archivo de SIMAT Detallado al blob storge de Azure está descrito en el Documento Técnico de SIMAT.

7.2.1.1. Procesos:

Para la integración de la fuente se desarrollan procesos de Extracción, Transformación y Carga.

7.2.1.2. Extracción:

Para realizar la extracción de los datos se configura el servicio de Azure Data Factory. Para que se mapee la información del archivo de texto DetalladoSimat.txt, se estructuren los campos que lo componen y guarde los datos en las columnas definidas dentro de la tabla de staging, conservando la misma estructura de la fuente.

255

Tabla 2: Estructura archivo detallado SIMAT.

ATRIBUTO	TIPO DATO	ATRIBUTO	TIPO DATO
ANO	varchar(50)	SISBEN_TRES	varchar(50)
ESTADO	varchar(50)	PER_ID	varchar(50)
JERARQUIA	varchar(50)	DOC	varchar(50)
INSTITUCION	varchar(255)	TIPODOC	varchar(100)
DANE	varchar(50)	APELLIDO1	varchar(100)
CALENDARIO	varchar(50)	APELLIDO2	varchar(100)
SECTOR	varchar(50)	NOMBRE1	varchar(100)
SEDE	varchar(255)	NOMBRE2	varchar(50)
CODIGO_DANE_SEDE	varchar(50)	GENERO	varchar(50)
CONSECUTIVO	varchar(50)	FECHA_NACIMIENTO	varchar(50)
ZONA_SEDE	varchar(50)	MATRICULACONTRATADA	varchar(100)
JORNADA	varchar(50)	FUENTE_RECURSOS	varchar(100)
GRADO_COD	varchar(50)	INTERNADO	varchar(100)
GRUPO	varchar(50)	EPS_ESTUDIANTE	varchar(255)
MODELO	varchar(255)	NUM_CONTRATO	varchar(50)
MOTIVO	varchar(255)	APOYO_ACADEMICO_ESPECIAL	varchar(255)
FECHAINI	varchar(50)	SRPA	varchar(255)
FECHAFIN	varchar(50)	DISCAPACIDAD	varchar(255)
NUI	varchar(50)	FECHA_INSERCION	datetime
ESTRATO	varchar(50)	CODIGO_CARGA	uniqueidentifier

7.2.1.3. Transformación:

En este proceso se aplican las reglas de negocio requeridas para la integración del archivo detallado simat con el repositorio de datos.

Únicamente se cargarán los registros de novedades de estudiantes que hayan sido previamente cargados en el proceso de integración de la Fuente SIMAT, para asegurar la integridad de los datos.

El estudiante debe pertenecer a un colegio y a una sede.

El registro del estudiante debe cumplir con criterios de completitud asociados a los datos maestros (grado, jornada).

256

No puede haber dos registros asociados a un estudiante con misma fecha y hora.

Los registros que no se cargan a la tabla destino de novedades de los estudiantes en el repositorio de datos, por no cumplir con cualquiera de las reglas anteriormente descritas, se almacenarán en tablas de staging, con la descripción del motivo de rechazo[19].

7.2.1.4. Carga:

En este proceso se tienen en cuenta únicamente los registros que cumplen con las reglas de negocio. Si se identifica que la novedad es nueva se insertará validando que los cambios que existan, comparando con la última novedad por estudiante y se marcará como activa.

7.2.1.5. Programación de ejecución de procesos

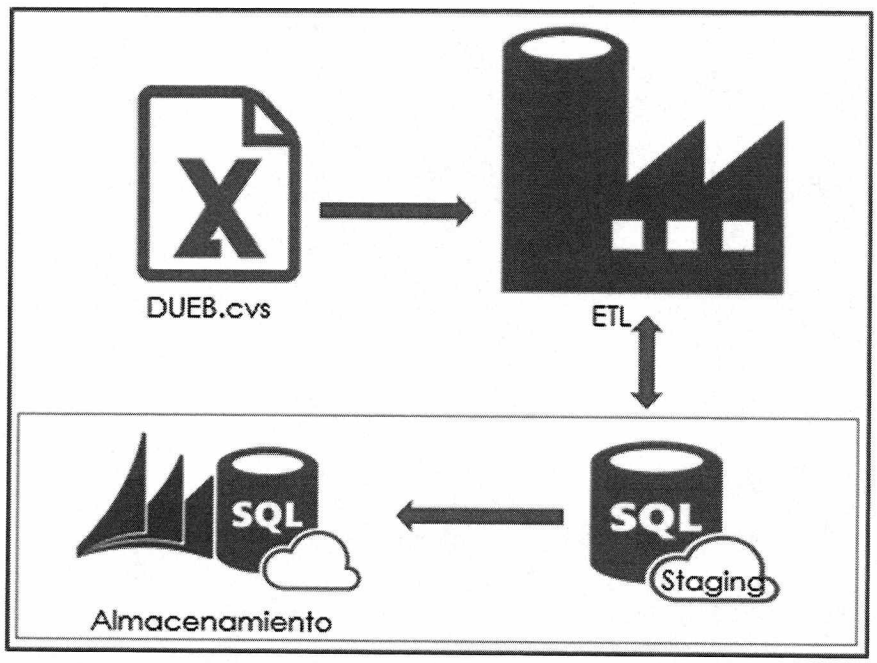
Para la integración de este proceso se requiere la creación de un flujo de trabajo dentro del servicio de Azure Data Factory, la programación de ejecución del proceso será diaria y posterior a la ejecución del proceso de integración de la fuente SIMAT.

Nota: El proceso depende de la ejecución de la integración de la fuente SIMAT en los siguientes casos: 1. Primer cargue del detallado. 2. Teniendo en cuenta la periodicidad de ejecución del cargue de SIMAT[20].

7.3. Integración Fuente DUEB.

257

[19] Los nombres de los procesos, las tablas de origen destino y rechazo, están especificados en el Documento técnico de SIMAT y en el Diccionario de datos.
[20] La programación de la ejecución o la ejecución manual de los procesos está documentada en el Manual Técnico Simat.



7.3.1. Definición de la fuente:

DUEB (Directorio Único de Establecimientos Bogotá: Archivo en formato csv., que contiene la información de las instituciones. Colegio, sede.

Nota: El procedimiento de carga del archivo de DUEB al blob storge de Azure está descrito en el Documento Técnico de XRM.

7.3.2. Procesos:

Para la integración de la fuente se desarrollan procesos de Extracción, Carga.

7.3.2.1. Extracción:

258

Para realizar la extracción de los datos se configura el servicio de Azure Data Factory. Para que se mapee la información del archivo DUEB.csv, se estructuren los campos que lo componen y guarde los datos en las columnas

definidas dentro de la tabla de staging, conservando la misma estructura de la fuente.

Tabla 3: Estructura DUEB.

ATRIBUTO	TIPO DATO	ATRIBUTO	TIPO DATO	ATRIBUTO	TIPO DATO
ID	varchar(MAX)	MODELO_JORNADA_UNICA_PRIMARIA	varchar(MAX)	GRADOS_JORNADA_NOCTURNA_SECUNDARIA	varchar(MAX)
LOCALIDAD	varchar(MAX)	GRADOS_JORNADA_UNICA_SECUNDARIA	varchar(MAX)	MODELO_JORNADA_NOCTURNA_SECUNDARIA	varchar(MAX)
NOMBRE_LOCALIDAD	varchar(MAX)	MODELO_JORNADA_UNICA_SECUNDARIA	varchar(MAX)	GRADOS_JORNADA_NOCTURNA_MEDIA	varchar(MAX)
DANE11_ESTABLECIMIENTO_EDUCATIVO	varchar(MAX)	GRADOS_JORNADA_UNICA_MEDIA	varchar(MAX)	MODELO_JORNADA_NOCTURNA_MEDIA	varchar(MAX)
DANE12_ESTABLECIMIENTO_EDUCATIVO	varchar(MAX)	MODELO_JORNADA_UNICA_MEDIA	varchar(MAX)	GRADOS_JORNADA_FIN_DE_SEMANA_PRIMARIA	varchar(MAX)
NOMBRE_ESTABLECIMIENTO_EDUCATIVO	varchar(MAX)	GRADOS_JORNADA_COMPLETA_PREESCOLAR	varchar(MAX)	MODELO_JORNADA_FIN_DE_SEMANA_PRIMARIA	varchar(MAX)
NIT_COLEGIO	varchar(MAX)	MODELO_JORNADA_COMPLETA_PREESCOLAR	varchar(MAX)	GRADOS_JORNADA_FIN_DE_SEMANA_SECUNDARIA	varchar(MAX)
DANE11_SEDE_EDUCATIVA	varchar(MAX)	GRADOS_JORNADA_COMPLETA_PRIMARIA	varchar(MAX)	MODELO_JORNADA_FIN_DE_SEMANA_SECUNDARIA	varchar(MAX)
DANE12_SEDE_EDUCATIVA	varchar(MAX)	MODELO_JORNADA_COMPLETA_PRIMARIA	varchar(MAX)	GRADOS_JORNADA_FIN_DE_SEMANA_MEDIA	varchar(MAX)
NOMBRE_SEDE_EDUCATIVA	varchar(MAX)	GRADOS_JORNADA_COMPLETA_SECUNDARIA	varchar(MAX)	MODELO_JORNADA_FIN_DE_SEMANA_MEDIA	varchar(MAX)
CONSECUTIVO_SEDE	varchar(MAX)	MODELO_JORNADA_COMPLETA_SECUNDARIA	varchar(MAX)	CARACTER_PARA_LA_MEDIA	varchar(MAX)
ORDEN_DE_SEDE	varchar(MAX)	GRADOS_JORNADA_COMPLETA_MEDIA	varchar(MAX)	ESPECIALIDAD_PARA_LA_MEDIA	varchar(MAX)
SECTOR	varchar(MAX)	MODELO_JORNADA_COMPLETA_MEDIA	varchar(MAX)	ENFASIS_PARA_EL_CARACTER_ACADEMICO_DE_LA_MEDIA	varchar(MAX)
CLASE	varchar(MAX)	GRADOS_JORNADA_MAÑANA_PREESCOLAR	varchar(MAX)	RECTOR_COLEGIOS	varchar(MAX)
NATURALEZA_JURIDICA	varchar(MAX)	MODELO_JORNADA_MAÑANA_PREESCOLAR	varchar(MAX)	TIPO_DE_DOCUMENTO_RECTOR	varchar(MAX)
ESTADO	varchar(MAX)	GRADOS_JORNADA_MAÑANA_PRIMARIA	varchar(MAX)	NUMERO_DE_DOCUMENTO_RECTOR	varchar(MAX)
ZONA	varchar(MAX)	MODELO_JORNADA_MAÑANA_PRIMARIA	varchar(MAX)	CARGO	varchar(MAX)
DIRECCION_GEOREFERENCIACION	varchar(MAX)	GRADOS_JORNADA_MAÑANA_SECUNDARIA	varchar(MAX)	DISCAPACIDAD_POR_CATEGORIA	varchar(MAX)
DIRECCION_COMUN_REPORTE_COLEGIOS	varchar(MAX)	MODELO_JORNADA_MAÑANA_SECUNDARIA	varchar(MAX)	TALENTOS_O_CAPACIDADES_EXCEPCIONALES	varchar(MAX)
BARRIO_GEOREFERENCIACION	varchar(MAX)	GRADOS_JORNADA_MAÑANA_MEDIA	varchar(MAX)	GRUPOS_ETNICOS_	varchar(MAX)
BARRIO_COMUN	varchar(MAX)	MODELO_JORNADA_MAÑANA_MEDIA	varchar(MAX)	SITUACION_DE_CONFLICTO_	varchar(MAX)
TELEFONO	varchar(MAX)	GRADOS_JORNADA_TARDE_PREESCOLAR	varchar(MAX)	BILINGÜE	varchar(MAX)
FAX	varchar(MAX)	MODELO_JORNADA_TARDE_PREESCOLAR	varchar(MAX)	REGIMEN_Y_CATEGORIA_DE_COSTOS	varchar(MAX)
EMAIL	varchar(MAX)	GRADOS_JORNADA_TARDE_PRIMARIA	varchar(MAX)	CHIP_1	varchar(MAX)
WEB	varchar(MAX)	MODELO_JORNADA_TARDE_PRIMARIA	varchar(MAX)	CPF	varchar(MAX)
CODIGO_POSTAL	varchar(MAX)	GRADOS_JORNADA_TARDE_SECUNDARIA	varchar(MAX)	ESTRATO_GEO	varchar(MAX)
CALENDARIO	varchar(MAX)	MODELO_JORNADA_TARDE_SECUNDARIA	varchar(MAX)	No_UPZ	varchar(MAX)
GENERO	varchar(MAX)	GRADOS_JORNADA_TARDE_MEDIA	varchar(MAX)	NOMBRE_UPZ	varchar(MAX)
GRADOS_JORNADA_UNICA_PREESCOLAR	varchar(MAX)	MODELO_JORNADA_TARDE_MEDIA	varchar(MAX)	SECTOR_CENSAL	varchar(MAX)
MODELO_JORNADA_UNICA_PREESCOLAR	varchar(MAX)	GRADOS_JORNADA_NOCTURNA_PRIMARIA	varchar(MAX)	X	varchar(MAX)
GRADOS_JORNADA_UNICA_PRIMARIA	varchar(MAX)	MODELO_JORNADA_NOCTURNA_PRIMARIA	varchar(MAX)	Y	varchar(MAX)
				OTRO	varchar(MAX)

Para asegurar el correcto funcionamiento de este proceso, esta estructura no debe ser modificada en el origen (archivo .csv), ni agregar, ni mover, ni quitar columnas, ni filas del título.

Los registros que al ser cargados en el staging tengan datos erróneos en el id, o que presenten problemas de calidad de datos en los dane12 del colegio y de la sede se rechazarán y se almacenaran en una tabla del mismo esquema, con la descripción del motivo de rechazo[21].

7.3.2.2. Carga:

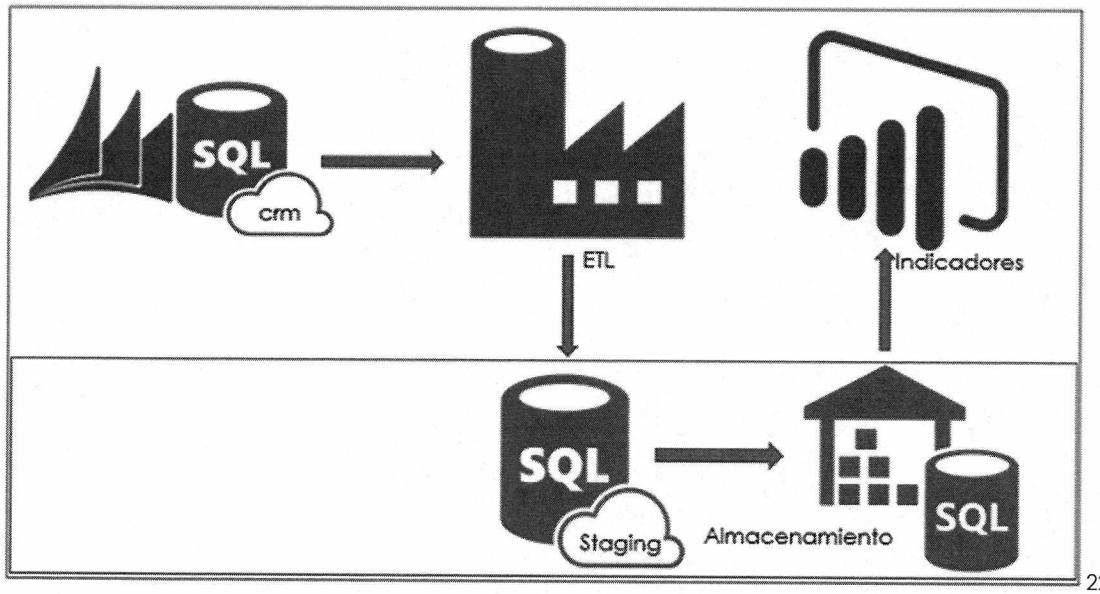
EL proceso de Cargue del Archivo DUEB a Dynamics 365/CRM, comprende el mapeo de los campos de la tabla staging con la tabla destino ACCOUNT, se utiliza el Proceso de UPSERT que tiene nativamente el conector con el fin de poder realizar el proceso de Actualización e inserción de datos.

[21] Los nombres de las tablas de origen destino y rechazo, están especificados en el Documento técnico de XRM y en el Diccionario de datos.

7.3.2.3. Programación de ejecución de procesos

Para la integración de este proceso se requiere la creación de un flujo de trabajo dentro del servicio de Azure Data Factory, al no tener una periodicidad definida la ejecución debe ser manual.

7.4. Integración Fuente CRM.



22

7.4.1. Definición de la fuente.

CRM: Sistema de información que contiene la digitalización de los procesos de la subsecretaria de accesos y permanencia, la dirección de servicios administrativos, instituciones, escalafón y la oficina de personal.

7.4.1.1. Extracción:

269

Para realizar la extracción de los datos se configura el servicio de Azure Data Factory, para que establezca una conexión con las entidades del Dynamics

[22] Los indicadores se describen de forma general en el numeral 7.6.

CRM, extraiga los datos requeridos para la integración y los almacene en las tablas de staging, conservando la misma estructura del origen, de acuerdo con los módulos que comprenden el CRM para la integración.

Tablas de Instituciones: Información relacionada con los colegios y sedes que prestan el servicio de educación en el distrito.

Modulo	Entidad
Instituciones	Account
Instituciones	Area
Instituciones	AreaEvaluacion
Instituciones	AreaLinea
Instituciones	Barrio
Instituciones	DatoAcademico
Instituciones	DiscapacidadSede
Instituciones	EscalaInsuficiencia
Instituciones	Especialidad
Instituciones	EspecialidadSede
Instituciones	EstratoGeoecon
Instituciones	EtniaSede
Instituciones	Evaluacion
Instituciones	Idioma
Instituciones	IdiomaSede
Instituciones	Localidad
Instituciones	Modelo
Instituciones	Motivo
Instituciones	NivelAcademico
Instituciones	ResultadoAcademico

Tablas de Acceso y permanencia: Asocia la información relacionada con el bienestar de los estudiantes para el acceso a la educación en el distrito (Alimentación, Movilidad, Infraestructura etc.).

261

Modulo	Entidad
Acceso y permanencia	BaseMaestra
Acceso y permanencia	BaseMaestraDetalle
Acceso y permanencia	BienestarAlimentacion
Acceso y permanencia	BienestarMovilidad
Acceso y permanencia	CategoriaBM
Acceso y permanencia	CodigoPlantaFisica
Acceso y permanencia	DotacionContrato
Acceso y permanencia	DotacionContratoElem
Acceso y permanencia	DotacionContratoElemDet
Acceso y permanencia	Elemento
Acceso y permanencia	EntregasAlimentacion
Acceso y permanencia	GrupoElemento
Acceso y permanencia	Instalacion
Acceso y permanencia	LineaProfundizacion
Acceso y permanencia	ModalidadAlimentacion
Acceso y permanencia	ModalidadInstalacion
Acceso y permanencia	PiezaRural
Acceso y permanencia	Predio

Tablas de Escalafón: Información de solicitudes de ascenso o inscripción de docentes del sector privado.

Modulo	Entidad
Escalafon	Anotaciones
Escalafon	AreaEnsenanza
Escalafon	Cargo
Escalafon	Contact
Escalafon	Escalafon

Tablas de Servicios Públicos: Facturas pagadas de los servicios públicos de las instalaciones educativas propias o arrendadas.

Modulo	Entidad
Servicios públicos	CDPFactura
Servicios públicos	ConceptoFactura
Servicios públicos	CuentaContratoFactura
Servicios públicos	Factura
Servicios públicos	InstalacionFactura
Servicios públicos	OperadorFactura
Servicios públicos	ProveedorFactura

262

7.4.1.2. Transformación:

En este proceso se aplican las reglas de negocio requeridas para la integración de la fuente con el repositorio de datos.

Instituciones

- Los colegios deben tener un único código dane12 de colegio.
- Las sedes deben tener un único dane12 de sede.
- No puede haber colegios o instituciones con los códigos dane vacíos o nulos.

Acceso y Permanencia

- Si se identifica que un registro es nuevo debe ser insertado en el destino.
- Si se identifica que un registro existente en el destino tiene algún cambio, debe ser actualizado.

Servicios Públicos

- Si se identifica que un registro es nuevo debe ser insertado en el destino.
- Si se identifica que un registro existente en el destino tiene algún cambio, debe ser actualizado.

Escalafón

- Se considera que el registro de un docente es único por número de identificación.

Los registros que no se cargan en las tablas destino de colegios, sedes y docentes, por no cumplir con cualquiera de las reglas que corresponden a su proceso se almacenan en tabla de staging con el motivo de rechazo[23].

263

[23] Los nombres de los procesos, las tablas de origen destino y rechazo, están especificados en el Documento técnico de XRM y en el Diccionario de datos.

7.4.1.3. Carga:

En este proceso se tienen en cuenta únicamente los registros que cumplen con las reglas de negocio. Si se identifican registros nuevos se insertarán en las tablas destino, si un registro tiene alguna diferencia se actualizará.

7.4.1.4. Programación de ejecución de procesos

Para la integración de este proceso se requiere la creación de varios flujos de trabajo dentro del servicio de Azure Data Factory. La periodicidad de ejecución de los procesos será diaria. El orden de ejecución estará dado de la siguiente manera[24]:

a) Primero se ejecutarán los procesos que integren entidades con datos maestros.

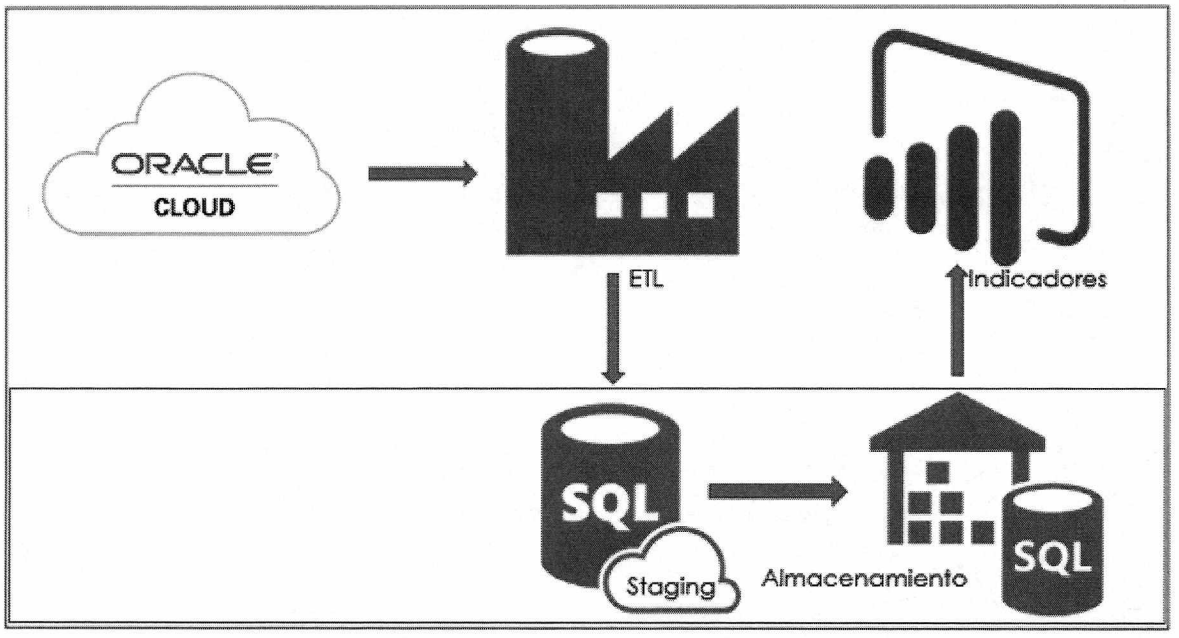
b) En segundo lugar, el proceso que contiene integra la información de instituciones.

c) Después irán los procesos de los módulos (acceso y permanencia, servicios públicos y escalafón).

264

[24] La programación de la ejecución o la ejecución manual de los procesos está documentada en el Manual Técnico XRM.

7.5. Integración Fuente Humano.



25

7.5.1. Definición de la fuente:

Humano: Humano es el sistema de talento humano de la Secretaría de Educación, el cual contiene la información de los docentes y personal administrativo del distrito; este sistema a dispuesto de una vista que contiene información básica del personal incluyendo la ubicación en colegios/sedes del distrito.

7.5.2. Procesos:

Para la integración de la fuente se desarrolla proceso de Extracción, Transformación y Carga.

265

[25] Los reportes se pueden construir a futuro, acorde a la necesidad del usuario.

7.5.2.1. Extracción

Para realizar la extracción de los datos se configura el servicio de Azure Data Factory. Para que establezca una conexión con la vista de Docentes dispuesta por el Sistema de Información Humano, extraiga los datos requeridos para la integración y los almacene, conservando la misma estructura de la fuente.

Tabla 4: Estructura de la vista de Humano.

Atributo	Tipo de Dato	Atributo	Tipo de Dato
TIPODEDOCUMENTO	varchar(255)	TELEFONODELDOCENTE	varchar(255)
NUMERODEDOCUMENTO	varchar(255)	EMAILDELDOCENTE	varchar(255)
PRIMERNOMBRE	varchar(255)	FECHAINGRESO	datetime
SEGUNDONOMBRE	varchar(255)	FECHARETIRO	datetime
PRIMERAPELLIDO	varchar(255)	ESTADO	varchar(255)
SEGUNDOAPELLIDO	varchar(255)	UBICACIONCOLEGIO	varchar(255)
GENERO	varchar(255)	CODIGODANECOLEGIO	varchar(255)
TIPODEFUNCIONARIO	decimal(18, 0)	DIRECIONCOLEGIO	varchar(255)
CARGO	varchar(255)	TELEFONOCOLEGIO	varchar(255)
TIPO	varchar(255)	EMAILCOLEGIO	varchar(255)
NUMVINCULACION	decimal(18, 0)	NOMBRESEDE	varchar(255)
ORIGENVINCULACION	decimal(18, 0)	CODIGODANESEDE	varchar(255)
FUENTEFINANCIANCION	varchar(255)	TELEFONOSEDE	varchar(255)
JORNADADELDOCENTE	varchar(255)	EMAILSEDE	varchar(255)
GRADOACTUALESCALAFON	varchar(255)	TIPOREG	decimal(18, 0)
REGIMEN1278O2277	varchar(255)	IDCARGA	uniqueidentifier
DIRECIONDELDOCENTE	varchar(255)	FECHAINSERCION	datetime

7.5.2.2. Transformación:

En este proceso se aplican las reglas de negocio requeridas para la integración de la fuente con el repositorio de datos.

Los números de documento que se tendrán en cuenta para la inserción del docente y las novedades del docente son los que en el atributo *tipo* tiene las siguientes descripciones. ('DIRECTIVO DOCENTE', 'DOCENTE').

Se insertará o actualizarán al docente con el registro que tiene la última fecha de ingreso.

Las novedades se insertarán por número de documento y fecha de ingreso, este proceso no tiene actualización.

Los registros que no se cargan en las tablas destino de docentes y novedades de docentes, por no cumplir con cualquiera de las reglas que corresponden a su proceso se almacenan en tabla de staging con el motivo de rechazo[26].

7.5.2.3. Carga:

En este proceso se tienen en cuenta únicamente los registros que cumplen con las reglas de negocio. Para registros nuevos inserción y para las existentes actualizaciones.

7.5.2.4. Programación de ejecución de procesos

Para la integración de este proceso se requiere la creación de un flujo de trabajo dentro del servicio de Azure Data Factory. La periodicidad de ejecución de este proceso será diaria.

Las actividades de este proceso estarán configuradas dentro del flujo de trabajo del módulo de escalfon del CRM[27].

26

[26] Los nombres de los procesos, las tablas de origen destino y rechazo, están especificados en el Documento técnico de Humano y en el Diccionario de datos.
[27] La programación de la ejecución o la ejecución manual de los procesos está documentada en el Manual Técnico Humano.

7.6. Indicadores

Para este proceso se crearán los siguientes reportes asociados a la vista de datos de instituciones y estudiantes[28].

- Ficha Matricula: Reporte con la matrícula de los estudiantes del distrito.
- Jornada Única: Reporte de la matrícula de los estudiantes que pertenecen a la jornada única.
- Primera Infancia: Reporte con la matrícula de los estudiantes de preescolar.
- Alimentación Escolar: Reporte que muestra las raciones de alimentos que se entregan a las instituciones para garantizar el acceso y la permanencia a la educación de los estudiantes del distrito, teniendo en cuenta diferentes variables.
- Construcciones – Mejoramientos: Reporte que muestra los proyectos de mejoramiento de la infraestructura de las instituciones. Para garantizar el acceso y la permanencia de la educación para los estudiantes del distrito.
- Construcciones – Proyectos: Reporte que muestra proyectos de construcción de la infraestructura de las instituciones. para garantizar el acceso y la permanencia de la educación para los estudiantes del distrito.
- Dotaciones: Reporte de la dotación que se le entrega las instituciones para garantizar el acceso y la permanencia de la educación para los estudiantes del distrito.
- Movilidad Escolar: Reporte donde se muestra los tipos de movilidad y la cantidad de estudiantes beneficiados con movilidad de las instituciones, para garantizar el acceso y la permanencia a la educación de los estudiantes del distrito, teniendo en cuenta diferentes variables.
- Servicios Públicos: Reporte que muestra los pagos de servicios públicos que realizan las instituciones.
- Cobertura: Reporte de la matricula por localidad, capacidad excepcional y discapacidad de los estudiantes del distrito.

8. Recomendaciones

- **Repositorio Datos (Servicio SQL Server):**

Creación de índices para que las búsquedas sean rápidas y eficientes.

Se requiere un usuario administrador del servicio de Azure SQL Server, para la creación de usuarios, índices, procedimientos almacenados, vistas etc.

Se recomienda crear usuarios con permisos de lectura para el acceso a objetos específicos del repositorio de datos. Así mismo para la integración del repositorio con otras aplicaciones,

Definir tiempos de retención de los registros rechazados en cada una de las integraciones.

- **Integración (Azure Data Factory)**

Para crear instancias de Data Factory, la cuenta de usuario que use para iniciar sesión en Azure debe ser un miembro de los roles colaborador o propietario, o de administrador de la suscripción de Azure.

Para crear y administrar recursos secundarios para Data Factory incluidos los conjuntos de datos, servicios vinculados, canalizaciones, desencadenadores y entornos de ejecución de integración, el usuario debe pertenecer al rol Colaborador de Data Factory en el nivel de grupo de recursos.

Mantener actualizada la versión del servicio Azure Data Factory.

269

[28] La explicación a detalle de los reportes se encontrará en el manual técnico Simat, manual técnico XRM.