Tesis de Maestría

DISEÑO DE UN TABLERO DE GESTIÓN COMO SOPORTE PARA LA TOMA DE DECISIONES EN EL AMBITO DE LA RECOLECCIÓN DE RESIDUOS EN LA CIUDAD AUTÓNOMA DE BUENOS AIRES

Lic. Daniel Alejandro Poblet

FACULTAD DE INGENIERÍA

Maestría en Tecnología de la Información

Director de Tesis

Prof. Dr. Nicolás D'Ippolito

Buenos Aires, Argentina 2018

APROBACIÓN DEL COMITÉ

_____
Tutor

_____
Presidente del jurado

_____
Jurado

_____
Jurado Externo

Defensa de la tesis
Buenos Aires, Capital Federal, A los ___ días del mes _____ del _____

A Julieta

la compañía perfecta en la vida y en cada nuevo emprendimiento.

A Joaquín

porque desde muy pequeño fuiste la motivación que me faltaba para poder terminar un camino que llevaba tiempo recorriendo.

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE SIGLAS

Sigla	Significado
ARIMA	AutorRegresivo Integrado de Promedio Móvil
BI	Inteligencia de Negocios
BSC	Cuadro de Mando Integral
BYOL	Traiga su propia licencia
CABA	Ciudad Autónoma de Buenos Aires
CIO	Chief Information Officer
CPU	Unidad Central de Procesamiento
DBMS	Sistemas de gestión de Bases de Datos
DESM	Alisamiento exponencial doble / Double Exponential Smoothing
DGEyC	Dirección General de Estadística y Censos
DW	Data Warehouse
ERP	Planificación de recursos empresariales
GCP	Google Cloud Platform
GIS	Sistema de información geográfica
IaaS	Infraestructura como Servicio
IoT	Internet de las cosas / Internet of things
KPI	Indicadores de Desempeño / Key Performance Indicator
ML	Machine Learning
MMDB	Sistema de Bases de Datos en Memoria
NIST	Instituto Nacional de Estándares y Tecnología
NORAD	Mando Norteamericano de Defensa Aeroespacial
OLAP	Sistemas de Procesamiento Analítico en Línea
OLTP	Sistemas de Procesamiento Transaccional en Línea
PaaS	Plataforma como Servicio
PAL	Librerías de Análisis Predictivo
PM	Gerente de Proyecto / Project Manager
SaaS	Software como Servicio
SAP	Sistemas, Aplicaciones y Productos
SESM	Alisamiento exponencial simple / Simple Exponential Smoothing
SUACI	Sistema Único de Atención Ciudadana
TDWI	The Data Warehousing Institute
TESM	Alisamiento exponencial Triple / Triple Exponential Smoothing
TI	Tecnología de Información
VM	Máquinas Virtuales
WOPR	Respuesta del Plan Operativo de Guerra

1. INTRODUCCIÓN

El promedio de recolección diaria de residuos por persona en la Ciudad Autónoma de Buenos Aires (CABA) durante el año 2016 fue de 1,35 Kg según los datos proporcionados por la Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA) sobre base de datos de CEAMSE (Estadísticas y Censos, 2016). Un valor mayor al de la región de Latino América que está estimado en 1.1 Kg (Hoornweg & Bhada-Tata, 2012, pág. 9). Los recursos naturales son transformados en bienes de consumo cuya vida útil es menor en relación con el tiempo necesario para su degradación (Odriozola, 2014), lo que obliga a pensar que los sistemas actuales, rellenos sanitarios y/o incineración deben ser reemplazados o mejorados. El continuo crecimiento de la población incrementa la necesidad de optimizar el sistema de recolección.

1.1 Definición Del Problema

El Gobierno de la Ciudad de Buenos Aires, dentro del marco de la LEY N.º 1854/05 sancionada el 24 de noviembre de 2005, propone que la Ciudad adopte, para la el tratamiento de los residuos sólidos urbanos, el concepto de Basura Cero (Legislatura de la Ciudad Autónoma de Buenos Aires, 2005), demostrando el compromiso del Gobierno de la Ciudad de Buenos Aires pone a este tema.

Este programa se orienta no solamente al tratamiento y el reciclaje, sino también al diseño de los productos de modo que tengan una vida útil más larga y se produzcan con materiales no tóxicos y reciclables.

La implementación de un programa como Basura Cero lleva años, donde se debe educar a la población, incrementar los controles y buscar herramientas que permitan mejorar, monitorear y tomar decisiones acertadas.

La Ciudad de Buenos Aires ha estado incorporando cambios de hábito en la población, como la separación de residuos en origen, según indica “La propuesta es que cada habitante de la ciudad clasifique los residuos antes de sacarlos a la calle, colocando los papeles y cartones en una bolsa verde que sea identificable por los cartoneros" (Como se cita en Odriozola, 2014) de esta forma los materiales pueden ser reutilizados o reciclados, mientras que el resto, siguen su proceso normal. Esta iniciativa se lanza en conjunto con el sistema de contenedores urbanos, Campanas y Puntos Verdes distribuidos en la Ciudad de Buenos Aires.

1.1.1 Ley N.º 1854/05 – Basura Cero

La LEY N.º 1854/05 (2005) tiene como objetivo establecer el conjunto de pautas, principios, obligaciones y responsabilidades para la gestión integral de los residuos sólidos urbanos que se generen en el ámbito territorial de la Ciudad Autónoma de Buenos Aires, en forma sanitaria y ambientalmente adecuadas, a fin de proteger el ambiente, seres vivos y bienes. En este sentido la Ciudad adopta como principio para la problemática de los residuos sólidos urbanos el concepto de Basura Cero.

En dicha norma se definió al concepto de Basura Cero (2005) como: “el principio de reducción progresiva de la disposición final de los residuos sólidos urbanos, con plazos y metas concretas, por medio de la adopción de un conjunto de medidas orientadas a la reducción en la generación de residuos, la separación selectiva, la recuperación y el reciclado” (Ley 1854/05, 2005, Art 2°).

La siguiente tabla muestra el cronograma de reducción de residuos, sobre una línea base de 1.497.656 toneladas, como fue reglamentado mediante el decreto Nro. 639/07 (2007).

También sostiene que las operaciones de gestión de residuos sólidos urbanos se deben realizar sin poner en riesgo la salud humana.

Tabla 1: Cronograma de Reducción de Residuos (2007)

Año	Porcentaje de Reducción	Cantidad Máxima para disponer en relleno Sanitario (Tn)
2010	30%	1.048.359
2012	50%	748.828
2017	75%	374.414
2020	100%	A definir por la Autoridad de Aplicación

Fuente: Adaptado del Decreto Nro. 639/07, Anexo I, Art: 6°.

1.1.2 Recolección de Residuos Sólidos.

El Gobierno de la Ciudad de Buenos Aires (2017) tiene un fuerte compromiso con la reducción de los residuos que se entierran en los rellenos sanitarios de CEAMSE. De las 6.000 toneladas de residuos promedio que genera la ciudad por día, aproximadamente 4.000 corresponden a residuos húmedos y 2.000 a residuos áridos o restos de obra. El servicio de recolección funciona en su mayor parte por la noche y se realizan aproximadamente 480 viajes en camiones compactadores que recolectan las 4.000 toneladas de residuos húmedos y los llevan a tres estaciones de transferencia.

Existen algunos tipos de residuos que no transitan los circuitos de los materiales reciclables ni la basura, y para los cuales la disposición debe realizarse de una manera específica.

Estos tipos de residuos pueden ser: Escombros, residuos voluminosos y restos de poda, cuyo retiro debe ser solicitado y se realiza de manera gratuita. El servicio se puede solicitar el servicio por medio del Sistema Único de Atención Ciudadana (SUACI) o mediante llamado al 147 para coordinar día y horario para la recolección.

Entre los servicios disponibles se encuentran:

- Retiro de residuos voluminosos: muebles, electrodomésticos, artefactos sanitarios, cerramientos, maderas, chatarras u otros residuos de gran volumen.
- Retiro de escombros: restos de obra, escombros, áridos, en bolsas adecuadas y hasta 500 kilos.
- Retiro de restos de poda: restos de plantas, ramas, malezas y residuos de jardín.

1.2 Justificación del Estudio

Davenport (2009) dice que una de las formas de mejorar la toma de decisiones es refinando el análisis de datos. Además, destaca que las organizaciones que utilizan el análisis de datos como ventaja competitiva aplican los datos de manera de optimizar operaciones en grados sin precedencia y transforman la tecnología de una herramienta de soporte en un arma estratégica.

1.3 Objetivos

1.3.1 Objetivo General

El objetivo de esta tesis es el de analizar la problemática de recolección de residuos en la Ciudad Autónoma de Buenos Aires (CABA) utilizando las diferentes fuentes de datos públicas, para diseñar un tablero de gestión que, incorporando técnicas

como Machine Learning, Business Analytics y Análisis de Información Geográfica (GIS de sus siglas en inglés Geographic information system), sirva como soporte para la toma de decisiones.

1.3.2 Objetivos Específicos

- Analizar la problemática de la recolección de residuos en la Ciudad Autónoma de Buenos Aires.
- Proponer un modelo analítico que permita evaluar los principales problemas encontrados.
- Unificar las diferentes fuentes de información pública disponibles:
    - Buenos Aires Open Data (https://data.buenosaires.gob.ar).
    - Estadísticas y Censos (https://www.estadisticaciudad.gob.ar)

  para alimentar el modelo diseñado.
- Utilizar técnicas de Machine Learning y Business Analytics para procesar, aprender y generar modelos predictivos en base a esta información.
- Incorporar análisis geográfico para procesar y presentar la información.
- Desarrollar una aplicación prototipo la cual permita aplicar el modelo propuesto para evaluar los resultados obtenidos.

2. FUNDAMENTOS TEÓRICOS

2.1 Tableros de Control

De acuerdo con Kaplan y Norton (1997), los tableros de control son un conjunto de indicadores que constituyen una herramienta de diagnóstico. Los tableros de control pueden cubrir todos los niveles de detalle de una empresa, desde el nivel más estratégico e integral hasta el nivel operativo alimentado directamente por los sistemas transaccionales, como ser el sistema de gestión empresarial o ERP. El tablero de control de más alto nivel será el denominado cuadro de mando, consultado a nivel directivo y construido a partir de la información del nivel de detalle inferior.

Kaplan y Norton (1997) comparan el cuadro de mando de una empresa con los tableros de navegación en un avión. Se necesita un destino claro (los objetivos estratégicos) y el monitoreo de múltiples medidas en simultáneo para asegurarse la llegada a destino, utilizando los indicadores como retroalimentación para poder realizar ajustes. La visibilidad de las medidas está dada por el tablero de control y el cuadro de mando Integral (abreviado en inglés como BSC, por Balanced Scorecard). De acuerdo con Kaplan y Norton, el BSC fue desarrollado para medir el desempeño de una organización desde las siguientes perspectivas:

- Financiera.
- Del cliente.
- De procesos internos.
- De aprendizaje y crecimiento.

El tablero de control debe mantener una alineación con la estrategia y objetivos de una organización. La relación con la estrategia es tan fuerte que, un cuadro de mando integral debe poder traducirse en un mapa estratégico, la representación visual de cómo

los objetivos estratégicos se traducen en objetivos específicos de cada área una de las perspectivas mencionadas anteriormente.

De acuerdo con Kaplan (1999) y a diferencia de las organizaciones con fines de lucro, las organizaciones gubernamentales utilizan la perspectiva financiera como referencia para control y restricción del gasto en lugar de como objetivo de ganancia. En otras palabras, las finanzas pierden relevancia en el contexto de las organizaciones de sector público, dando lugar a la medición del éxito según la eficiencia y eficacia con la que se alcanzan las metas de la organización. Por consiguiente, existe una diferencia fundamental entre los cuadros de mando integral aplicados en el sector público respecto de su aplicación en organizaciones en el sector privado. Debido a esta diferencia se trata de comenzar por la perspectiva del cliente en el sector público, en lugar de por la perspectiva financiera. Este cambio surge como consecuencia de entender que la perspectiva financiera en las organizaciones sin fines de lucro funciona como una restricción en lugar de como un objetivo. Al ser la perspectiva financiera solo una restricción, no pueden colocarse objetivos medibles a los que aspirar.

Una vez establecidos los objetivos estratégicos de alto nivel, se procede a generar objetivos a mayor nivel de detalle, hasta alcanzar objetivos que cada miembro de la organización pueda monitorear. De acuerdo con Kaplan (1999), la comunicación de los cuadros de mando integrales a todos los miembros de una organización juega un papel clave en la generación de compromiso por parte de cada individuo. Así como la estrategia se establece “de arriba hacia abajo”, la medición de resultados provendrá, principalmente, de la información transaccional. Se concluye, que las transacciones individuales, originadas en las acciones cotidianas de los niveles operativos, serán de vital importancia para alcanzar las metas y objetivos que sustenten la estrategia.

2.2 Business Intelligence (BI) y Business Analytics (BA)

Uno de los activos más importantes en una organización es la información. Este activo es generalmente utilizado de dos maneras: almacenar los registros operativos y toma de decisiones.

El término Business Intelligence (BI) estuvo dentro de los principales temas en la agenda de los CIOs, el concepto de BI no es algo nuevo y no hay una única definición que describa a esta herramienta.

En su investigación sobre la historia de los sistemas de soporte de decisión, Power (2007) relata que la primera definición de BI data del año 1989, hecha por Howard Dresner, también reconocido como el padre de BI, y como miembro de Gartner Research, define como: "El conjunto de software y soluciones para recopilar, consolidar, analizar y proporcionar acceso a los datos de manera de permitir a los usuarios empresariales tomar mejores decisiones negocio".

Otros autores afirman que la primera definición data de 1958 y fue elaborada por H. P. Luhn en un artículo titulado "A Business Intelligence System" presentando un concepto muy similar al utilizado en la actualidad. (Como se cita en Gibson, Arnott, & Jagielska, 2004 y Chee, y otros, 2009)

Al mismo tiempo, otra definición que es altamente reconocida es la proporcionada por The Data Warehousing Institute (TDWI[1]) define BI como "El proceso, tecnologías y herramientas necesarias para convertir datos en información, información en conocimiento y conocimiento en planes que dirijan acciones de negocio rentables. BI abarca Data Warehouse, herramientas analíticas y gestión de conocimiento"

[1] TDWI es el principal instituto de educación en el área de BI y Data Warehousing, fundado en 1995. Se dedica a la investigación, educación y certificación de profesionales dentro del sector de tecnología de la información.

Davenport Y Harris (2007) definen BA como una subcategoría dentro de BI, basada en estadísticas, predicción y optimización. Además, consideran a BI más enfocado hacia las capacidades de reporte con respecto a BA.

Según Gartner (s.f.) BA está compuesta por soluciones utilizadas para construir modelos de análisis y simulaciones que permiten crear escenarios, comprendiendo la realidad para predecir estados futuros. Está compuesto por: minería de datos, análisis predictivo, análisis aplicado y estadístico. Siendo una aplicación disponible para los usuarios de negocio que contiene, a menudo, escenarios preconfigurados para la industria.

En síntesis, podemos entender el BI como las formas de recolectar y entender datos del pasado, mientras que el BA nos permite construir una visión más clara del futuro. Ambas herramientas pueden complementarse para elaborar un análisis detallado del funcionamiento y futuro de una empresa, con el fin de tomar mejores decisiones.

2.2.1 Modelos OLTP y OLAP

La necesidad de analizar la información y convertirla en conocimiento hace necesario el tratamiento diferencial para grandes volúmenes de información, lo que nos permite distinguir 2 tipos de sistemas: Operacionales y Analíticos.

A mediados de 1970, los sistemas de procesamiento transaccional en línea OLTP (On Line Transaction Processing), hicieron posible un acceso más rápido a los datos, abriendo un nuevo campo para los negocios y el procesamiento de la información. Las computadoras comenzaron a utilizarse para tareas que no habían sido capaces en el pasado, sin esté cambio, si el mundo se hubiera quedado usando de archivos en cinta magnética, la mayoría de los sistemas que hoy damos por sentado, no serían posibles. (Inmon, 2002)

Los usuarios de un sistema transaccional mueven la rueda de la organización, Crean las órdenes, registran nuevos clientes, monitorean el estado de las actividades operativas y registran las quejas entre otras actividades. Procesan una transacción por vez, ejecutando los procesos de la empresa reiteradamente, no almacenan el historial, aunque mantienen actualizado el estado actual de la organización representando toda la información transaccional que es generada en el accionar diario, además de las fuentes externas de las que puede disponer:

- Archivos de textos
- Hipertextos
- Hojas de cálculos
- Informes
- Bases de datos transaccionales

El procesamiento analítico en línea OLAP (On Line Analytic Processing), es el motor de consultas especializado de una Base de Datos. Las herramientas OLAP, son una tecnología de software para análisis en línea, administración y ejecución de consultas, que permiten inferir información del comportamiento del negocio.

Su principal objetivo es el de brindar respuestas rápidas a preguntas complejas, para interpretar la situación del negocio y tomar decisiones. Cabe destacar que lo que es realmente interesante en OLAP, no es la ejecución de consultas tradicionales, sino la posibilidad de utilizar herramientas avanzadas que permiten navegar en las diferentes dimensiones, para explotar profundamente la información.

En contraste con un sistema transaccional, los usuarios de un sistema analítico miran las ruedas de la organización, evalúan su desempeño, comparando la información actual con la histórica, analizan los nuevos clientes registrados y los motivos de las quejas recibidas. Tienen interés en que los procesos de negocio trabajen adecuadamente,

aunque nunca tratan con una transacción a la vez. Analizan la información, corrigen el rumbo, marcan el camino y definen la velocidad con que las ruedas de la organización deben ir para alcanzar el destino.

Las diferencias entre un sistema transaccional y uno analítico están bien marcadas y en algunos casos podrían hasta ser opuestas, la siguiente tabla muestra las principales diferencias entre los aspectos fundamentales de cada uno:

Tabla 2: Diferencias entre sistemas transaccionales y analíticos.

Característica	Transaccional	Analítico
Origen de Datos	Transacciones diarias, un único origen	Diversos orígenes de datos
Contenido de Datos	Los datos son actuales	Datos históricos, derivados, totalizados
Objetivo de los datos	Operación Diaria	Decisiones
Tipo de Acceso	Lectura, escritura, actualización y borrado	Sólo Lectura
Uso	Predecible, repetitiva	Aleatoria, para situaciones específicas, para descubrimiento
Tiempo de Respuesta	Rápido (Segundos)	Dependiendo de la complejidad de la consulta (Minutos)
Cantidad de Usuarios	Dependiendo del sistema (miles)	Menor cantidad de usuarios (Cientos)
Diseño de la Base de datos	Modelo entidad relación. Datos normalizados	Modelo dimensional. Datos desnormalizados
Requerimiento de Espacio	Menor cantidad de espacio. Archivado frecuentemente	Mayor cantidad de espacio ya que almacena datos históricos (una vez archivados del transaccional)
Consultas	Sencillas, preferentemente para consultar un número pequeño de registros	Complejas, requiere agregaciones de datos y múltiples dimensiones
Frecuencia de Acceso	Frecuente. Cubre las operaciones del día a día	Menor frecuencia, bajo demanda.

Información expuesta	Información de los procesos de negocio	Vista multidimensional de las actividades del negocio

Fuente: adaptado de A Critical Comparison Between Distributed Database Approach and Data Warehousing, por Hossain, Sohrab; Islam, Farhana; Karim, Razuan; Siddique, Kazy Noor-E-Alam , p.5.

2.2.2 Data Warehouse

Según Kimball y Ross (2013), el objetivo de Data Warehouse y Business Intelligence (DW/BI) puede ser desarrollado caminando por los pasillos de la organización y escuchando a los gerentes del negocio, sus preocupaciones se transforman en los fundamentos de DW/BI:

- Debe lograr que la información sea de fácil acceso.
- Debe presentar información consistente.
- Se debe adaptar a los cambios de la organización.
- Debe presentar la información a tiempo.
- Debe proteger la información como principal activo de la organización.

Una iniciativa de Business Intelligence (BI) requiere de una estructura de Data Warehouse (DW). La presentación de datos debe cumplir 2 requerimientos:

- Muestre los datos en un formato amigable al usuario final
- Provea un buen rendimiento en las consultas.

Para ello el modelado dimensional es la técnica mayormente utilizada.

Kimball y Ross (2013) sostienen que, aunque los modelos dimensionales a menudo se desarrollan en bases de datos relacionales, son bastante diferentes de los modelos de tercera forma normal, que se enfocan en eliminar la redundancia de datos. En general los modelos dimensionales se implementan utilizando diagramas de tipo

estrella, nombrados así por la semejanza de su representación con la estructura de una estrella.

La estructura de un diagrama estrella se compone de 2 componentes principales:

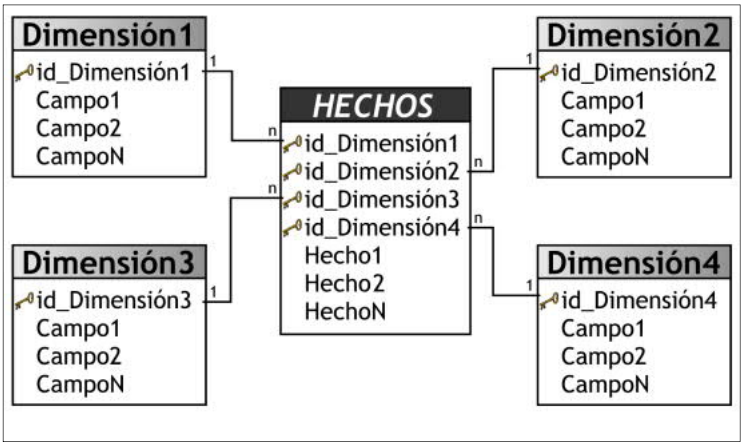
Tabla de Hechos

La tabla de hechos en un modelo dimensional almacena el resultado de las mediciones de rendimiento de eventos de procesos de negocios de una organización.

Tabla de Dimensiones

Contiene el detalle del contexto asociado con un evento del negocio. Complementa la información detallada en una tabla de hechos.

Figura 1: Tabla de Hechos y Dimensiones

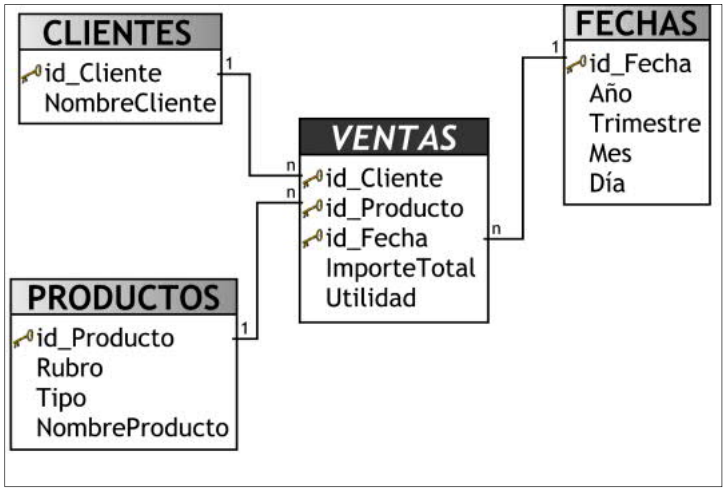


Fuente: de DATA WAREHOUSING: Investigación y Sistematización de Conceptos, por Bernabeu, Ricardo Dario, p.43.

Esquema Estrella

El esquema en estrella consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves. En la siguiente ilustración se puede apreciar un esquema en estrella estándar:

Figura 2: Ejemplo de Diagrama Estrella

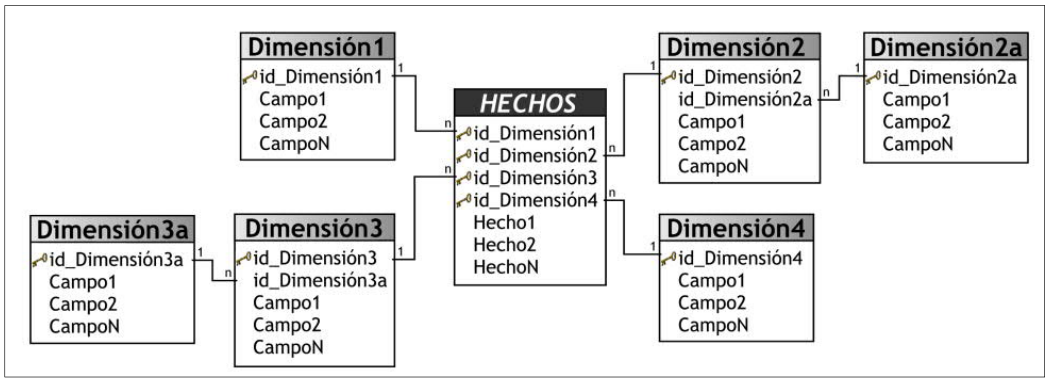


Fuente: de DATA WAREHOUSING: Investigación y Sistematización de Conceptos, por Bernabeu, Ricardo Dario, p.50.

Esquema Copo de Nieve

Este esquema representa una extensión del modelo en estrella cuando las tablas de dimensiones se organizan en jerarquías de dimensiones.

Figura 3: Esquema Copo de nieve



Fuente: de DATA WAREHOUSING: Investigación y Sistematización de Conceptos, por Bernabeu, Ricardo Dario, p.51.

2.3 Machine Learning

Según Mitchell (1997) nos encontramos frente a un programa de Machine Learning (ML) cuando dicho programa aprende de la experiencia **E** con respecto a alguna clase de tareas **T** y la medida de rendimiento **P**, si su rendimiento en tareas en **T**, medido por **P**, mejora con experiencia **E**.

Por lo tanto, para tener un caso de uso de ML se deben identificar estas tres características:

La clase de tarea (T): Describe la tarea que el algoritmo debe ejecutar o resolver.

La medida de rendimiento (P): se trata de un indicador que muestra el grado de cumplimiento de la clase de tarea luego de ejecutado el algoritmo, el cual debe mejorarse en la medida que aumenta la experiencia.

El origen de la experiencia (E): determina la forma en que el algoritmo va adquiriendo mayor experiencia.

Los problemas en los que ML es aplicable son variados, pero deben poder se expresados mediante estas tres características para garantizar su aplicabilidad. Un ejemplo práctico de esta aplicación es el diseño de un programa que aprenda a jugar a las damas.

Juego de Damas:

- Tarea **(T)**: Jugar a las Damas
- Medida de Rendimiento **(P)**: Porcentaje de partidos ganados contra los oponentes.
- Experiencia **(E)**: realizar juegos de práctica contra sí mismo.

ML no es un concepto nuevo, puede verse presente en la película Juegos de Guerra (WarGames) de 1983 (Wikipedia, s.f.) donde el sistema WOPR (War Operative Plan Response) es activado para jugar a “La Guerra Mundial Termonuclear” mediante

un acceso no autorizado al sistema y pensando que se trataba de un juego, pero el juego se torna tan real, que amenaza con realmente desatar una guerra mundial. Si planteamos este ejemplo en las características anteriormente descriptas, podemos decir:

Guerra Mundial termonuclear:

- Tarea **(T)**: Jugar a Guerra
- Medida de Rendimiento **(P)**: Porcentaje de peleas ganadas contra los oponentes.
- Experiencia **(E)**: realizar juegos de práctica contra sí mismo.

En la película y como resultado del aprendizaje, el sistema determina que no hay un escenario donde haya un ganador, por lo que concluye que la mejor jugada es no jugar.

Si bien aún no es posible que las computadoras aprendan de la misma forma que los humanos, existen algoritmos refinados que pueden ser utilizados para resolver algunos de los problemas cotidianos o de uso comercial que, como en el ejemplo anterior, pueden dar resultados no triviales y ayudar a encontrar nuevas soluciones a nuestros problemas.

Existen según Seeger (2002) dos tipos de escenarios aplicables a Machine Learning que pueden ser fácilmente identificados:

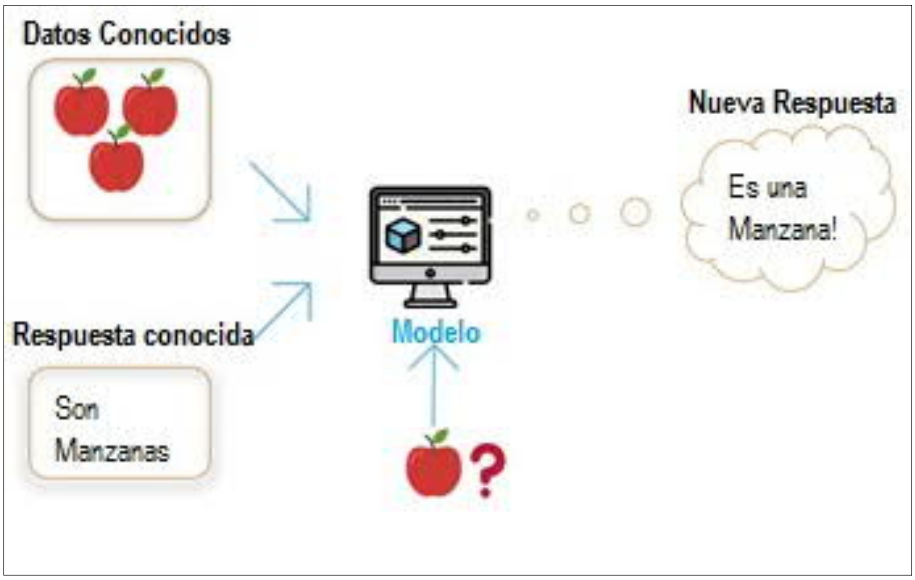
- Supervisado: Aprender con un maestro.
- No Supervisado: auto aprendizaje.

En estos dos tipos de escenarios, el principal diferencial surge de los datos utilizados para el entrenamiento. En el caso del entrenamiento supervisado el algoritmo se entrena utilizando datos ya clasificados, mientras que en el entrenamiento no supervisado el algoritmo no requiere datos hayan sido clasificados con anterioridad.

2.3.1 Entrenamiento Supervisado:

El modelo de entrenamiento supervisado, también conocido como modelo predictivo, requiere que los datos de entrenamiento se encuentren clasificados, de esta forma el algoritmo utilizado construye un modelo de predicción en base a un universo conocido, permitiendo predecir el resultado cuando se reciben nuevos datos.

Figura 4: Entrenamiento Supervisado

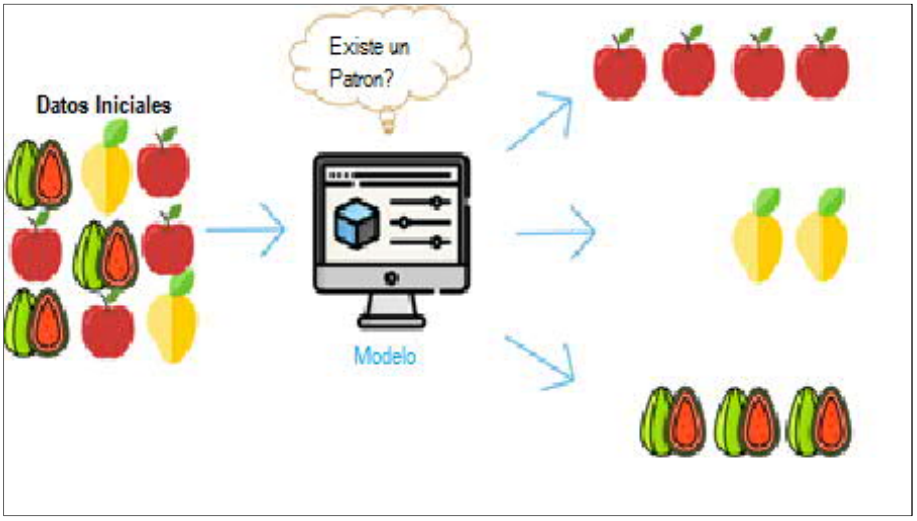


Fuente: Adaptado de What is the difference between supervised and unsupervised learning algorithms? Disponible en https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms, recuperado el 4 de agosto 2018.

2.3.2 Entrenamiento No Supervisado

También conocido como análisis descriptivo, este tipo de entrenamiento no requiere que los datos estén clasificados con anterioridad, sino que se encarga de buscar similitudes y regularidades en los datos de entrenamiento, con el fin de reconocer patrones entre ellos y generar un modelo en base a esos patrones. De esta forma, puede utilizar el modelo generado para describir un nuevo conjunto de datos.

Figura 5: Entrenamiento no supervisado

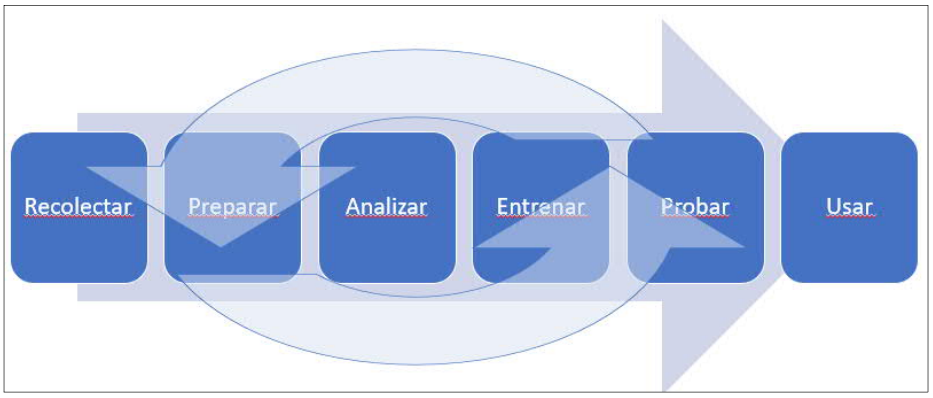


Fuente: Adaptado de What is the difference between supervised and unsupervised learning algorithms? Disponible en https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms, recuperado el 4 de agosto 2018.

2.3.3 Pasos Para El Desarrollo De Una Aplicación Con Machine Learning

Harrington (2012) propone un enfoque para desarrollar una aplicación con Machine Learning, el mismo consta de 6 pasos:

Figura 6: Pasos para desarrollar una aplicación de ML



Fuente: Elaboración propia.

1) Recolectar Datos: este paso consiste en recolectar la información, puede provenir de diferentes medios u orígenes, obtenida en tiempo real, o sencillamente utilizar un ser te datos ya existentes.

2) Preparar los Datos: Una vez obtenidos los datos, se deben preparar para que sean utilizables por un algoritmo de Machine Learning, si se define un formato estándar, se puede utilizar varios orígenes de datos.

3) Analizar los datos: Analizar los datos permite determinar la calidad de los mismos, determinar si la recolección se realiza de forma correcta, si hay muchos datos en blanco o revisar patrones obvios, se pueden utilizar herramientas para analizar 2, 3 o más dimensiones para visualizar la información.

4) Entrenar el Modelo: En este paso es donde se comienza a utilizar Machine Learning. En este paso se utilizan los algoritmos seleccionados para obtener un modelo. Este paso no es necesario si se utiliza entrenamiento no supervisado.

5) Probar el Modelo: Aquí es donde el modelo generado en el paso anterior se utiliza, en el caso de entrenamiento supervisado, se puede comprobar con los datos ya clasificados, en el caso de entrenamiento no supervisado, existen métricas para evaluar el nivel de certeza.

6) Usar el Modelo: En este paso se comienza a usar el modelo de forma automática para poder generar predicciones o resultados.

Durante cualquiera de estos pasos, se pueden presentar la necesidad de volver a un paso anterior para mejorar la recolección, calidad de datos, entrenar el algoritmo, hasta obtener el nivel de precisión deseado.

2.3.4 Series de Tiempo

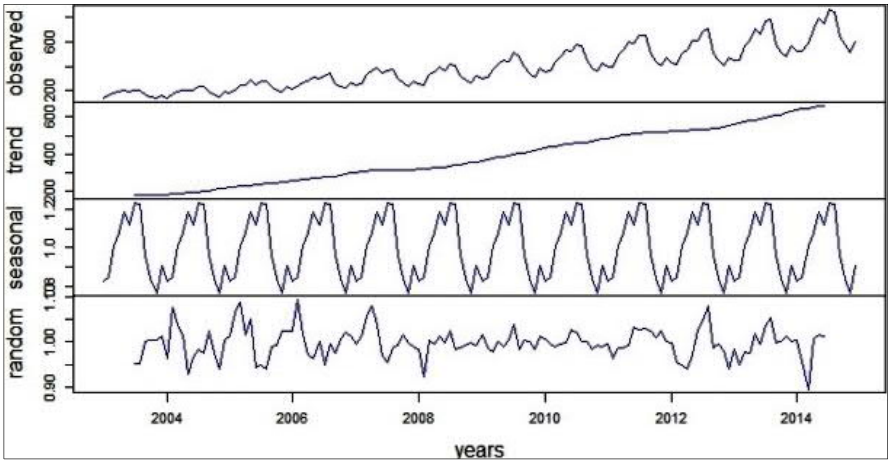
Las series de tiempo representan el resultado de la observación de los valores de una variable a lo largo del tiempo. Según Hamilton (1994) una serie de tiempo es una colección de observaciones ordenadas por la fecha de cada observación, con un inicio y fin determinados.

Una serie de tiempo se puede representar con un gráfico de secuencia, esto es, representar gráficamente cada valor de la observación X_t frente a un instante t, uniendo cada uno de los valores con segmentos.

El gráfico de secuencia nos permitirá observar cómo evoluciona la serie a lo largo del tiempo.

La predicción mediante series de tiempo (Kalekar, 2004) asume que la serie está formada por una combinación de patrones, más un error aleatorio. El objetivo es el de separar los patrones del error, mediante el entendimiento de la tendencia y su estacionalidad.

Figura 7: Descomposición de una serie de tiempo



Fuente: de Time Series Decomposition – Manufacturing Case Study Example (Part 2) por Roopam Upadhyay, disponible en https://i1.wp.com/ucanalytics.com/blogs/wp-content/uploads/2015/05/Time-Series-Decomposition-Plot.jpeg?w=608 , recuperado el 12 de agosto de 2018.

En general características de la serie de tiempo son:

Tendencia

Dirección general de la serie, es decir, hacia arriba, hacia abajo. Muestra el comportamiento que tendrá a largo plazo.

Estacionalidad

Indica el comportamiento periódico de la serie, la misma puede presentar patrones mensuales, trimestrales o anuales.

Ciclos

Representa los ciclos comerciales que pueden darse a largo plazo, no siempre está presente en una serie.

Valor residual

Ruido aleatorio que queda después de la extracción de todos los componentes.

2.3.5 Modelo ARIMA

Un modelo autorregresivo integrado de promedio móvil o ARIMA (acrónimo del inglés AutoRegressive Integrated Moving Average) es un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Se trata de un modelo dinámico de series temporales, es decir, las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes (Wikipedia, s.f.).

ARIMA (Upadhyay, s.f.) es una combinación de 3 partes, es decir, AR (AutoRegresivo), I (Integrado) y MA (Promedio Móvil). Una notación conveniente para el modelo ARIMA es ARIMA (p, d, q). Dónde p, d y q son los niveles para cada una de las partes AR, I y MA. Cada una de estas tres partes es un esfuerzo para hacer

que los residuos finales muestren un patrón de ruido blanco (o ningún patrón en absoluto). En cada paso del modelado ARIMA, los datos de series de tiempo pasan a través de estas 3 partes. La secuencia de tres pasos para el análisis ARIMA es la siguiente:

Integrated (I)

Resta series temporales con sus series anteriores para extraer tendencias de los datos.

AutoRegresivo (AR)

Extrae la influencia de los valores de los períodos anteriores en el período actual.

Moving Average (MA)

Extrae la influencia de los términos de error del período anterior sobre el error del período actual.

Aunque el modelo es poderoso, es difícil establecer el orden correcto para la ejecución de las secuencias, la función AUTOARIMA (SAP SE, s.f.) determina automáticamente el orden de los pasos del modelo ARIMA. Esta función es útil cuando se desconoce el valor y orden de los parámetros p, d y q. La función AUTOARIMA evalúa las posibles combinaciones de estas variables y retorna el mejor modelo encontrado.

2.3.6 Exponential Smoothing

Exponential Smoothing (Kalekar, 2004) es un procedimiento para revisar continuamente un pronóstico a la luz de la experiencia más reciente. Asigna pesos decrecientes de forma exponencial a medida que la observación sea más antigua. En otras palabras, a las observaciones recientes se les da relativamente más peso en los pronósticos que a las observaciones anteriores.

Single Exponential Smoothing (SESM)

Se utiliza para el pronóstico a corto plazo, generalmente sólo un mes en el futuro. El modelo asume que los datos fluctúan alrededor de una media razonablemente estable (sin tendencia o patrón de crecimiento consistente).

Double Exponential Smoothing (DESM)

Este método se utiliza cuando los datos muestran una tendencia. Funciona de forma muy parecida al método simple, salvo que se deben actualizar dos componentes en cada período: nivel y tendencia. El nivel es una estimación ajustada del valor de los datos al final de cada período. La tendencia es una estimación ajustada del crecimiento promedio al final de cada período.

Triple Exponential Smoothing (TESM)

Este método se utiliza cuando los datos muestran tendencia y estacionalidad. Para manejar la estacionalidad, tenemos que añadir un tercer parámetro. El conjunto de ecuaciones resultante se denomina método "Holt-Winters" (HW) en honor a los nombres de los inventores. Dependiendo del tipo de estacionalidad se puede dividir en dos modelos: Modelo Estacional Multiplicativo, utilizado cuando los datos muestran estacionalidad multiplicativa, y Modelo Estacional Aditivo, utilizado cuando los datos muestran estacionalidad aditiva.

Auto Exponential Smoothing

Dentro de las funciones provistas por la plataforma SAP HANA, existe la función Auto Exponential Smoothing esta función está diseñada para calcular los parámetros óptimos de las funciones Single, Double y Triple Exponential Smoothing detalladas anteriormente.

Esta función también produce los resultados de pronóstico basados en estos parámetros óptimos. Esta optimización se calcula explorando el universo de parámetros

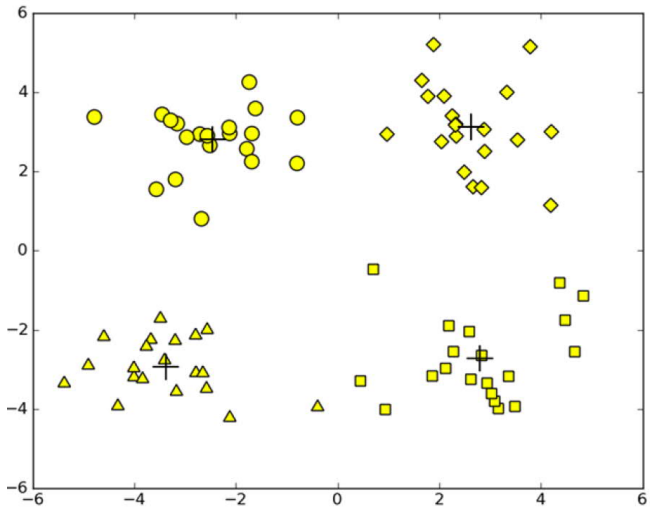
que incluye todas las combinaciones de parámetros posibles. La evaluación de la calidad se realiza comparando valores históricos y de pronóstico. Para evaluar la calidad de los parámetros se utilizan los indicadores MSE (error cuadrático medio) o MAPE (error porcentual absoluto medio).

Para evaluar la flexibilidad de la función, se lleva a cabo un esquema de entrenamiento y prueba. En otras palabras, se permite una partición de la serie temporal, de la cual la primera se utiliza para entrenar los parámetros, mientras que la segunda se aplica a la prueba.

2.3.7 K-Means

Este algoritmo recibe como parámetro un número K de grupos que deben encontrarse en el conjunto de datos. Para ello selecciona de forma aleatoria los centroides que representan cada uno de los grupos y los reubica, mediante el cálculo de la media con los todos miembros de un mismo grupo, iterativamente mientras la distribución cambie. Cuando los centroides no cambian de posición, el proceso termina.

Figura 8: Grupos resultantes luego de aplicar K-Means



Fuente: de Machine Learning in Action, por Peter Harrington, 2012, p. 212, recuperado el 15 de agosto del 2018.

2.4 Computación en la Nube

Según el Instituto Nacional De Estándares Y Tecnología (NIST) (2011), la computación en la nube es “un modelo para permitir el acceso a la red ubicuo, conveniente y bajo demanda a una red compartida grupo de recursos informáticos configurables (por ejemplo, redes, servidores, almacenamiento, aplicaciones y servicios) que puede aprovisionarse y liberarse rápidamente con un mínimo esfuerzo de gestión o interacción del proveedor de servicios.”

Asimismo, CISCO (2009) define a la computación en la nube como: “Recursos y servicios de Tecnología de Información (TI) que se abstraen de la infraestructura subyacente y se proporcionan a pedido y escala en un entorno multiusuario.” Y hace hincapié en tres atributos clave que debe poseer:

Bajo Demanda

Los recursos se pueden aprovisionar de inmediato cuando sea necesario y se liberan cuando ya no son requeridos, sólo es facturado cuando se usa.

A Escala

Significa que el servicio proporciona la ilusión de la disponibilidad de recursos infinita para cumplir cualquiera de los requerimientos de TI.

Entorno De Múltiples Propietarios

Los recursos se proporcionan a muchos consumidores desde un único punto de implementación, ahorrando al proveedor costos significativos.

La computación en la nube se presenta a menudo como una tecnología nueva, pero no lo es. Resulta más apropiado presentarlo como un nuevo enfoque o evolución que combina tecnologías ya conocidas (Rimal, Jukan, Katsaros, & Goeleven, 2010) como computación distribuida, computación en Malla, computación de utilidad, virtualización y clústeres de servidores entre otras.

Este modelo de nube se compone de cinco características esenciales, tres modelos de servicio y cuatro modelos de implementación.

2.4.1 Características Esenciales

Auto Servicio A Petición

Un consumidor puede aprovisionar unilateralmente capacidades informáticas, tales como tiempo de servidor y almacenamiento de red, según sea necesario, de forma automática y sin necesidad de interacción con el proveedor de servicios.

Amplio Acceso A La Red

Las capacidades están disponibles a través de la red y se accede mediante mecanismos estándar lo que promueve el uso desde cualquier dispositivo (Ej. teléfonos móviles, tabletas, computadoras portátiles y estaciones de trabajo).

Puesta En Común De Recursos

Los recursos informáticos del proveedor se agrupan para servir a múltiples consumidores utilizando un modelo multi-tenant, con diferentes recursos físicos y virtuales asignados dinámicamente y reasignados de acuerdo con la demanda del consumidor. Hay un sentido de independencia de ubicación en el que el cliente generalmente no tiene control o conocimiento sobre la ubicación de los recursos proporcionados, pero puede ser capaz de especificar la ubicación en un nivel superior de abstracción (por ejemplo, país, estado o centro de datos). Los ejemplos de recursos incluyen almacenamiento, procesamiento, memoria, ancho de banda de red y máquinas virtuales.

Rápida Elasticidad

Las capacidades se pueden aprovisionar y liberar elásticamente, en algunos casos automáticamente, permitiendo escalar o disminuir rápidamente en proporción a

la demanda. Para el consumidor las capacidades disponibles para aprovisionamiento a menudo parecen ser ilimitadas y pueden ser adquiridas en cantidades variables, en cualquier momento.

Servicio Medido

Los sistemas en la nube controlan y optimizan automáticamente el uso de los recursos gracias al aprovechamiento de la capacidad de medición en cierto nivel de abstracción apropiado para el tipo de servicio (Ej. almacenamiento, procesamiento, ancho de banda y cuentas de usuario activas). El uso de recursos puede ser monitoreado, controlado e informado, proporcionando transparencia para el proveedor y consumidor del servicio utilizado.

2.4.2 Modelo de Servicios

El término Computación en la Nube engloba tres niveles de prestación de servicio (Qi Zhang, Lu Cheng, & Raouf Boutaba, 2010)

Infraestructura Como Servicio (IaaS)

La capacidad provista al consumidor es la de aprovisionar procesamiento, almacenamiento, redes y otros recursos computacionales fundamentales para los cuales el consumidor es capaz de desplegarlos y ejecutar arbitrariamente software, que puede incluir sistemas operativos y aplicaciones.

Plataforma Como Servicio (PaaS)

Es la encapsulación de una abstracción de un ambiente de desarrollo y el empaquetamiento de una serie de módulos o complementos que proporcionan, normalmente, una funcionalidad horizontal (persistencia de datos, autenticación, mensajería, etc.).

De esta forma, un arquetipo de plataforma como servicio podría consistir en un entorno conteniendo una pila básica de sistemas, componentes o APIs preconfiguradas y listas para integrarse sobre una tecnología concreta de desarrollo.

Software Como Servicio (SaaS)

Es el más conocido de los tres modelos y el que suele tener como objetivo al cliente final, aquel que utiliza el software para ayudar, mejorar o cubrir algunos de los procesos de su empresa. El SaaS es aquella aplicación “consumida” a través de Internet, casi siempre a través del navegador, y donde tanto la lógica de la aplicación como los datos residen en la plataforma del proveedor.

2.4.3 Modelos de Implementación

Nube Privada

La infraestructura de la nube se aprovisiona para uso exclusivo de una sola organización que comprende múltiples consumidores (por ejemplo, unidades comerciales). Puede ser propiedad, administrado y operado por la organización, un tercero o una combinación de ellos, y puede existir dentro o fuera de las instalaciones de la organización.

Nube Comunitaria

La infraestructura de la nube se aprovisiona para uso exclusivo por parte de una comunidad de consumidores de organizaciones que comparten un mismo interés o propósito (Ej. misión, seguridad, política, etc.). Puede ser propiedad, administrado, y operado por una o más de las organizaciones en la comunidad, un tercero, o una combinación de ellos, y puede existir dentro o fuera de las instalaciones de la organización.

Nube Publica

La infraestructura de la nube está provista para uso abierto por el público en general. Puede ser propiedad, administrado y operado por una organización empresarial, académica o gubernamental, o alguna combinación de ellos. Existe en las instalaciones del proveedor de la nube.

Nube Hibrida

La infraestructura de la nube es una composición de dos o más nubes de distinta infraestructura (privadas, comunitarias o públicas) que siguen siendo entidades únicas, pero están vinculadas juntos por una tecnología estandarizada o propia que permite la portabilidad de datos y aplicaciones.

2.5 Bases de Datos

En su libro Fundamentos de Bases de Datos (Silberschatz, Korth, & Sudarshan, 2002) se define a una Base de Datos como un sistema de bases de datos es una colección de archivos interrelacionados y un conjunto de programas que permitan a los usuarios acceder y modificar estos archivos. Uno de los propósitos principales de un sistema de bases de datos es proporcionar a los usuarios una visión abstracta de los datos. Es decir, el sistema esconde ciertos detalles de cómo se almacenan y mantienen los datos.

Bajo la estructura de la base de datos se encuentra el modelo de datos: una colección de herramientas conceptuales para describir los datos, las relaciones, la semántica y las restricciones de consistencia. Los modelos de datos pueden ser del tipo modelo entidad-relación y el modelo relacional.

2.5.1 Modelo Entidad Relación (E-R)

El modelo de datos entidad-relación (E-R) está basado en una percepción del mundo real que consta de una colección de objetos básicos, llamados entidades, y de relaciones entre estos objetos. Las entidades se describen en una base de datos mediante un conjunto de atributos. Una relación es una asociación entre varias entidades.

La estructura lógica general de una base de datos se puede expresar gráficamente mediante un diagrama E-R, que consta de los siguientes componentes:

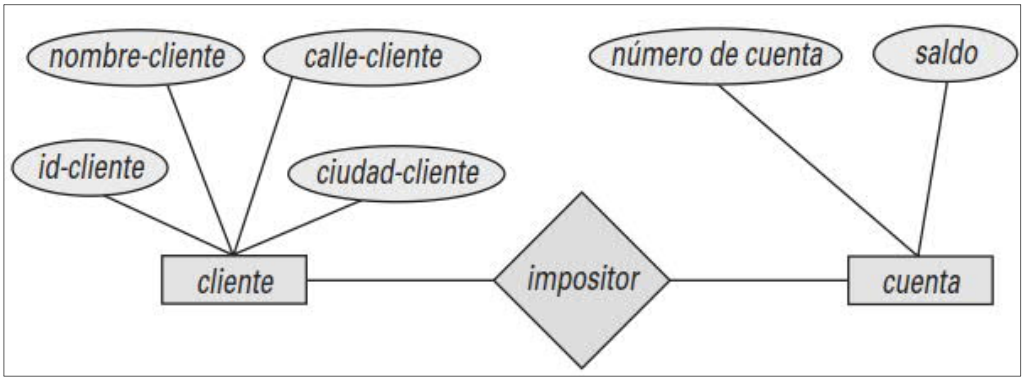
• Rectángulos, que representan conjuntos de entidades.

• Elipses, que representan atributos.

• Rombos, que representan relaciones entre conjuntos de entidades.

• Líneas, que unen los atributos con los conjuntos de entidades y los conjuntos de entidades con las relaciones.

Figura 9: Ejemplo diagrama E-R



Fuente: de Fundamentos de Bases de Datos, por Silberschatz, Korth, & Sudarshan, 2002, Cuarta Edición, p.6.

2.5.2 Modelo Relacional

En el modelo relacional se utiliza un grupo de tablas para representar los datos y las relaciones entre ellos. Cada tabla está compuesta por varias columnas, y cada columna tiene un nombre único.

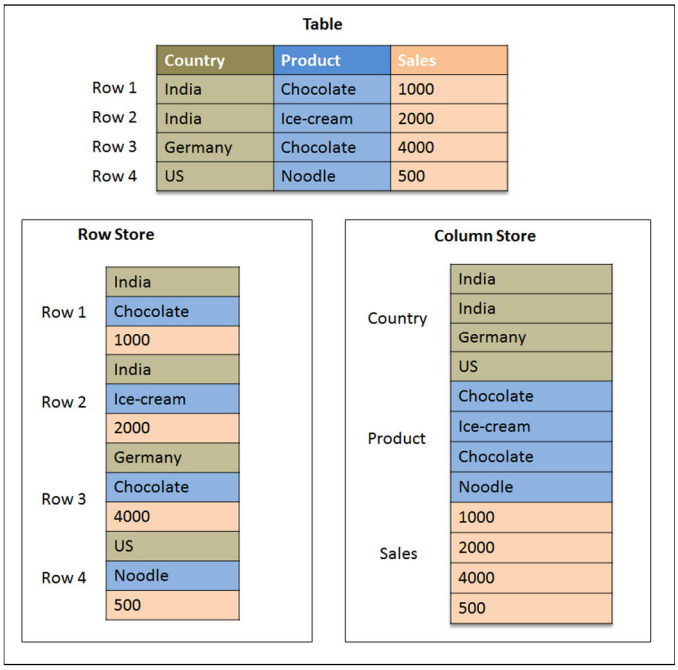
El modelo relacional es un ejemplo de un modelo basado en registros. Los modelos basados en registros se denominan así porque la base de datos se estructura en registros de formato fijo de varios tipos. Cada tabla contiene registros de un tipo particular. Cada tipo de registro define un número fijo de campos, o atributos. Las columnas de la tabla corresponden a los atributos del tipo de registro.

El modelo de datos relacional es el modelo de datos más ampliamente usado, y una amplia mayoría de sistemas de bases de datos actuales se basan en el modelo relacional.

2.5.3 Almacenamiento Columnar de Tablas (Row Column Store)

Debido a los avances en la velocidad de procesamiento de los CPU modernos (Boncz, Manegold, & Kersten, 1999) predijeron que el siguiente cuello de botella en el procesamiento se encontraba en el acceso a memoria. Al analizar los 3 aspectos de la memoria: Ancho de Banda, Latencia y Traducción de Direcciones (administración de memoria), que afectan la performance, notaron que la solución empleada por los fabricantes fue la de incorporar memoria cache en el subsistema de memoria, lo que genera una arquitectura más compleja que continúa siendo limitada y al menos que se tenga un tratamiento especial, el CPU podría estar un 95% del tiempo esperando acceder a memoria.

Figura 10: Comparación Almacenamiento Fila Vs Columnas



Fuente: de Column Vs Row Data Storage, disponible en http://saphanatutorial.com/column-data-storage-and-row-data-storage-sap-hana/ , recuperado el 10 de Julio 2018.

Por consiguiente, recomiendan utilizar estructuras de datos desglosada verticalmente. Esta organización permite la optimización de la lectura de datos mediante el uso de algoritmos que permiten aprovechar la arquitectura de los procesadores para paralelizar la búsqueda de la información.

Plattner (2009) también manifiesta que el almacenamiento organizado por columnas es el más adecuado para los procesadores actuales.

Los datos almacenados en columnas tienen mayor capacidad de compresión que los datos almacenados en filas. Los algoritmos de compresión funcionan mejor en datos con baja entropía de información (Abadi, Madden, & Hachem, 2008). Adicionalmente se produce un incremento significativo en la performance al permitir utilizar algoritmos

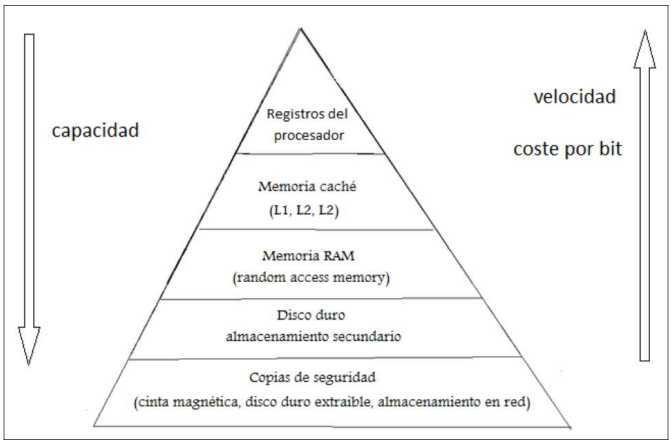
para procesar la información sin necesidad de descomprimirla (Abadi, Madden, & Ferreira, 2006) .

Según Plattner (2009) en un experimento realizado en base a los 5 años de datos históricos de una compañía cervecera Alemana con un total de 34 millones de registros almacenados en la tabla (1 millón de registros de esta tabla tiene un equivalente de 1 GB de espacio aproximado en disco) cuyo tamaño era de 35 GB fue comprimido en 8 GB de datos en su tabla columnar equivalente gracias a la compresión que este tipo de almacenamiento puede proveer.

2.5.4 Bases De Datos En Memoria (In Memory Database)

En un sistema de bases de datos en memoria (MMDB) los datos permanecen en memoria permanentemente. (Garcia Molina, 1992) A diferencia de los sistemas de Bases de Datos convencionales, los datos residen en disco y son cargados en memoria para ser accedidos por el motor de Bases de Datos.

Figura 11: Diagrama piramidal de Jerarquía de la Memoria



Fuente Jerarquía de Memoria, disponible en https://es.wikipedia.org/wiki/Jerarqu%C3%ADa_de_memoria , recuperado el 11 de Julio de 2018.

En el caso de los MMDB, los datos residen en disco a efectos del resguardo. Si bien ambos tipos de Bases de Datos tienen copias en disco y en memoria, la diferencia se basa en que en los MMDB la copia primaria se encuentra en memoria de forma permanente.

Ventajas

Según Artho (2014) las MMDB brindan las siguientes ventajas respecto a las BD tradicionales de almacenamiento en Disco.

- **Acceso Más Rápido A Los Datos**

  Dado que los mismos se encuentran en la memoria, se puede realizar un acceso aleatorio sin variar los tiempos de respuesta.

- **Mayor Poder De Cálculo**

  Cálculos sobre gran cantidad de información se pueden realizar en menor tiempo.

- **Múltiple Escritura**

  Una MMDB no tiene la restricción de una única escritura por página, ya que no existe el concepto de paginado.

- **Mejor Aprovechamiento Del CPU**

  Las MMDB están diseñadas para utilizar algoritmos que aprovechan mejor la estructura de los procesadores, lo que brinda un máximo de poder de procesamiento y cálculo.

Plattner (2014) también agrega una ventaja adicional y es que las MMDB permiten también simplificar la estructura de las BD mediante la eliminación de columnas agregadas y el uso de BD columnares. Plattner propone la eliminación de las columnas agregadas en el esquema de BD para dejar a la MMDB realizar el cálculo en tiempo de ejecución, aprovechando de esta forma su velocidad de procesamiento.

Desventajas

Al mismo tiempo Artho (2014) resalta las siguientes desventajas:

- **Almacenamiento De Datos Es Volátil**

Ya que los datos se almacenan en memoria RAM.

- **Tiempo De Respuesta Afectado Por La Latencia De Red**

Debido a la arquitectura de la BD en memoria, los sistemas que utilicen la información deben estar instalados en otro servidor, lo que genera que un tiempo de latencia adicional.

- **Cantidad De Información Disponible Es Limitada Por La Memoria Física**

Al ser la memoria RAM un recurso limitado, la cantidad de datos que pueden almacenarse se ve limitado por este factor. Teniendo el motor que usar procedimientos de intercambio a disco.

2.5.5 Plataforma SAP HANA

SAP define SAP HANA como "la plataforma de computación in-memory que le permite acelerar los procesos de negocio, brindar más Business Intelligence y simplificar su entorno de TI. Ofreciendo los cimientos para todas sus necesidades de datos, SAP HANA elimina la carga de tener que mantener sistemas heredados separados y datos en silos, lo que le permite operar en vivo y tomar mejores decisiones de negocio en la nueva economía digital (SAP SE, s.f.)".

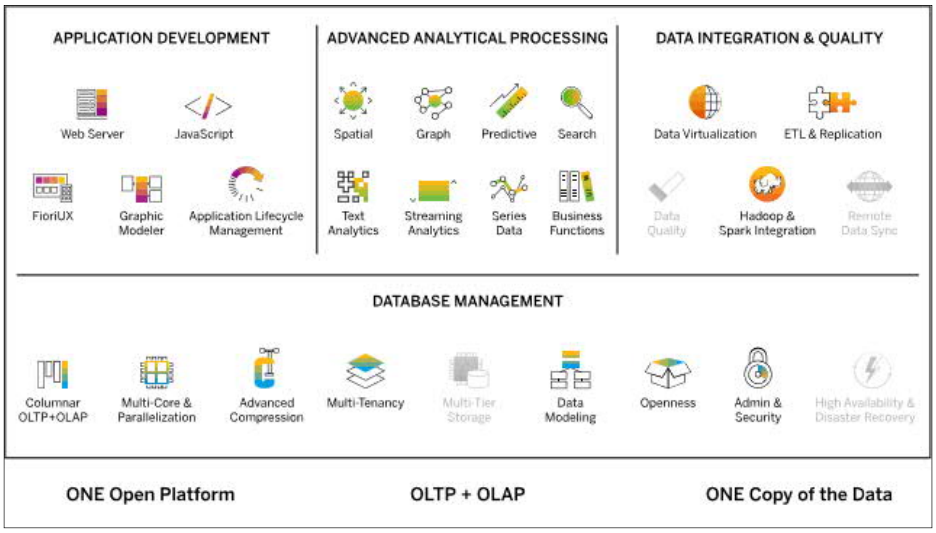
SAP HANA es la plataforma en que se basan la nueva generación de sistemas SAP tanto los que se encuentran on premise como en la nube.

Existe una versión liviana de SAP HANA, conocida como SAP HANA Express Edition (SAP SE, 2017). Esta versión puede ser instalada en una computadora personal,

o un servidor en la nube sin un costo de licencias si se utilizan hasta 32 GB de memoria, lo que lo hace muy interesante para la prueba y el desarrollo de prototipos.

SAP HANA Express Edition posee las mismas características analíticas y de aplicación que las versiones de SAP HANA empresariales, pero sin las capacidades de integración y alta disponibilidad. Puede ser migrada a una versión empresarial de ser necesario.

Figura 12: Arquitectura de SAP HANA Express Edition



Fuente: de Official SAP HANA, express edition tutorials and resources, disponible en https://www.sap.com/developer/topics/sap-hana-express.html, Recuperado el 17 de agosto de 2018.

SAP HANA combina dos características que la diferencian de otras plataformas: el uso de bases de datos en memoria y almacenamiento columnar de tablas.

2.5.6 Librería De Servicios Predictivos

La plataforma SAP HANA Express Edition posee de forma integrada una serie de librerías de predicción analítica y ML. Las Librerías PAL, por sus siglas en Ingles

Predictive Analysis Library, pueden ser utilizadas e invocadas mediante SQLScript, que es una versión extendida del SQL disponible en el producto.

Las librerías predictivas se pueden diferenciar en diferentes tipos, según su función:

Figura 13: Algoritmos PAL – HANA 2.0 SP3

Fuente: Elaboración propia.

Agrupación

Permite encontrar grupos afines dentro de los datos. Ej. ¿Qué personas de una muestra son candidatos para comprar un producto?

Clasificación

Permiten predecir una respuesta binaria. Ej. ¿Es una transacción fraudulenta o no?

Regresión

Permite predecir o clasificar un valor que no es binario. Ej.: Determinar el factor de riesgo de un conductor que solicita un seguro.

Asociación

Permite encontrar atributos que afectan una dimensión en particular. Ej. ¿Cuáles son los indicadores que determinan fallos futuros en los equipos

Serie de Tiempo

Permite predecir valores futuros en base a valores observados en el pasado. Ej. ¿Qué tan probable son las cancelaciones de vuelo en meses de invierno y en meses de verano?

Pre-Proceso

Los registros en la base de datos generalmente no están listos para el análisis predictivo por diversas razones: vienen en grandes cantidades, que pueden exceder la capacidad de un algoritmo, contienen observaciones ruidosas que pueden dañar la precisión de un algoritmo o algunos atributos están mal escalados, lo que puede hacer que un algoritmo sea inestable.

Estadística

Contiene aquellas funciones o técnicas que son utilizadas en análisis estadístico de datos sin procesar.

Análisis de Redes sociales

El análisis de redes sociales investiga la estructura de los grafos y la relación entre los nodos. Ej. Analizar los vínculos de amistad entre un conjunto de usuarios de Facebook.

Sistemas de Recomendación

El sistema de recomendación analiza los patrones de interés del usuario sobre los productos y proporciona recomendaciones personalizadas que se adaptan a los gustos del cada usuario. Ej. Recomendar una película en base a la calificación dada a otras películas.

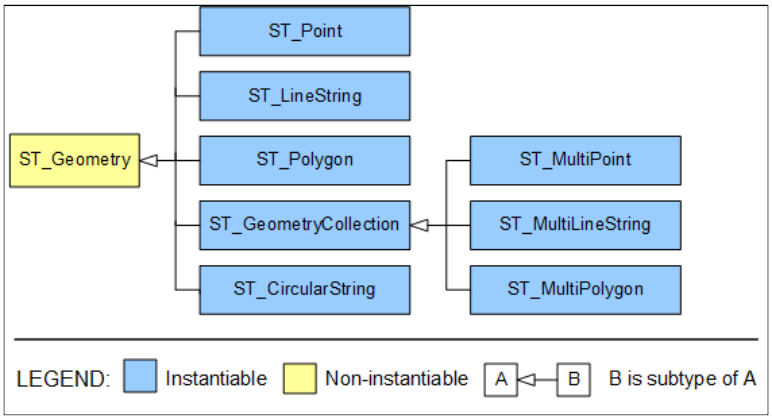
2.5.7 Referencia Espacial

Los datos espaciales (SAP SE, s.f.) son atributos que describen la posición, forma y orientación de los objetos en un espacio definido. Son representados como geometrías en 2 dimensiones, en forma de puntos, cadenas de líneas y polígonos.

SAP HANA incluye dentro de su funcionalidad un motor para administrar la referencia espacial de los objetos, para ello permite almacenar la información en sus tablas utilizando un tipo de dato espacial jerárquico: ST_Geometry dentro de una base de datos columnar.

La jerarquía de los tipos de datos espaciales está definida como muestra la siguiente figura. Todos los tipos de datos provienen del tipo de datos primitivo ST_Geometry.

Figura 14: Jerarquía de tipos de datos espaciales de SAP HANA



Fuente SAP HANA SPATIAL Reference Spatial type hierarchy, disponible en https://help.sap.com/viewer/cbbbfc20871e4559abfd45a78ad58c02/2.0.02/en-US/7a2ef60e787c10148e86fd0f4c60cb29.html , recuperado el 10 de agosto 2018.

SAP HANA provee los métodos para operar con los diferentes tipos de objetos geométricos. Estos métodos permiten calcular el área de una forma geométrica,

intersecciones entre figuras, distancias entre dos objetos, analizar si existe solapamiento, unión, etc.

Agrupación Espacial

SAP HANA proporciona agrupación espacial usando los algoritmos GRID, K-MEANS y DBSCAN. La agrupación espacial se puede realizar en un conjunto de puntos geoespaciales que estén almacenados en el sistema SAP HANA.

La agrupación espacial busca afinidades de una serie de objetos por los criterios definidos y genera grupos con características similares. Cada agrupación es un subconjunto de los datos originales.

Los criterios de agrupación dependen del algoritmo que se aplica en los objetos y en la parametrización del algoritmo.

3. DESARROLLO DE LA SOLUCION

3.1 Análisis De La Problemática De Recolección De Residuos

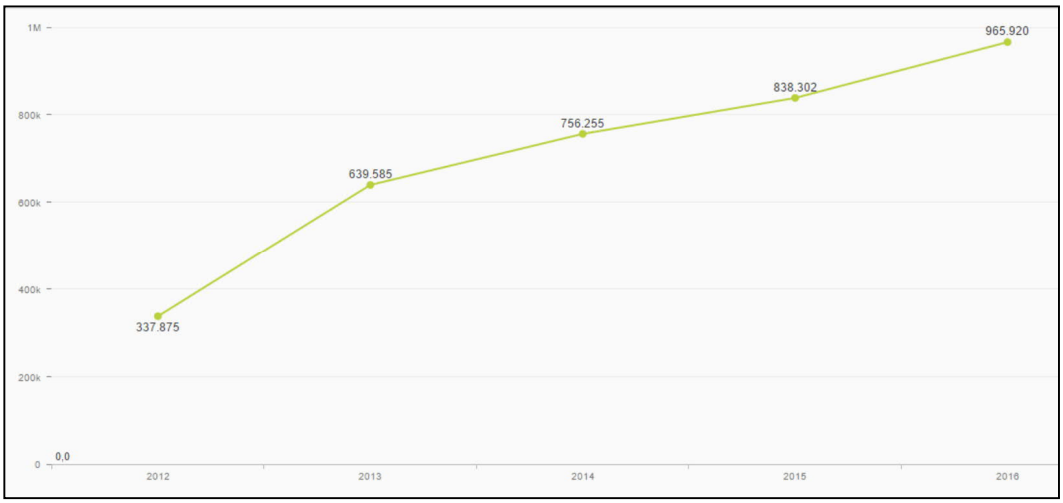
El objetivo de Basura Cero planteado para el año 2020 muestra sin duda un compromiso e interés por parte del Gobierno de la Ciudad de Buenos Aires muy grande, aunque no deja de ser ambicioso y difícil de lograr.

Muchas organizaciones del sector público (Kaplan R. , 1999) encuentran dificultades para desarrollar medidas apropiadas para la perspectiva financiera dentro de sus tableros de control, por lo que recomienda enfocarlo en la estrategia de la organización.

El objetivo del tablero de control es el de presentar información relevante que permita la toma de decisiones enfocado a dicha estrategia.

El Gobierno de la Ciudad de Buenos Aires, puso a disposición el sistema de Atención y Gestión Ciudadana (SUACI), mediante el cual permite a la población realizar sus solicitudes, quejas o reclamos dentro de diferentes rubros.

Figura 15: Participación Ciudadana por año



Fuente: Elaboración propia.

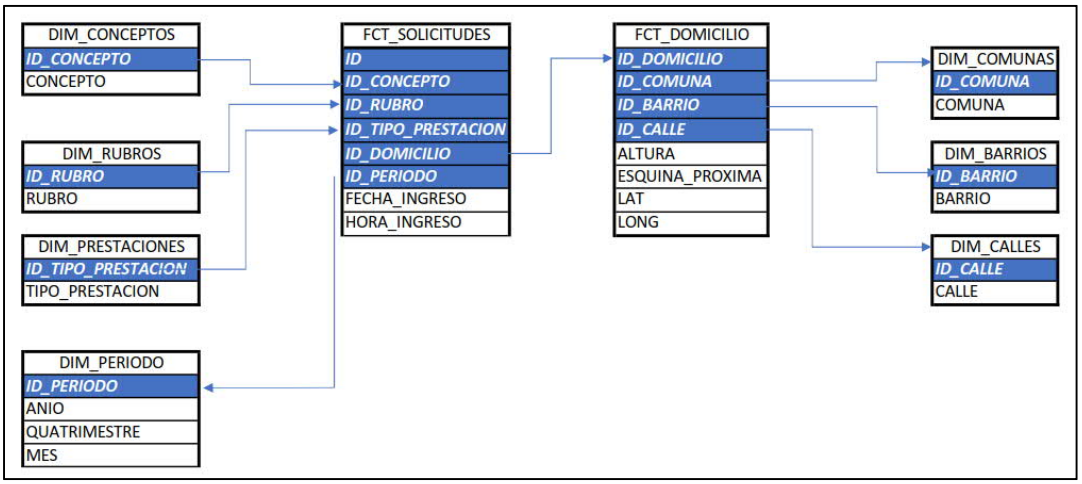
El acceso al SUACI se puede hacer a través de la página del Gobierno de la Ciudad de Buenos Aires [2] o llamando al 147. La figura 17 muestra la evolución de las solicitudes generadas al SUACI entre los años 2012 al 2016.

Los habitantes de la Ciudad de Buenos Aires muestran un continuo aumento en la participación y en la gestión, esto se ve reflejado en la cantidad de solicitudes ingresadas al sistema.

3.1.1 Modelo Analítico

Para el análisis de la información se utilizó un modelo Copo de Nieve, con la unión de 2 tablas de Hechos correspondientes a las Solicitudes realizadas y a sus Domicilios correspondientes:

Figura 16: Modelo Analítico propuesto



Fuente: Elaboración propia.

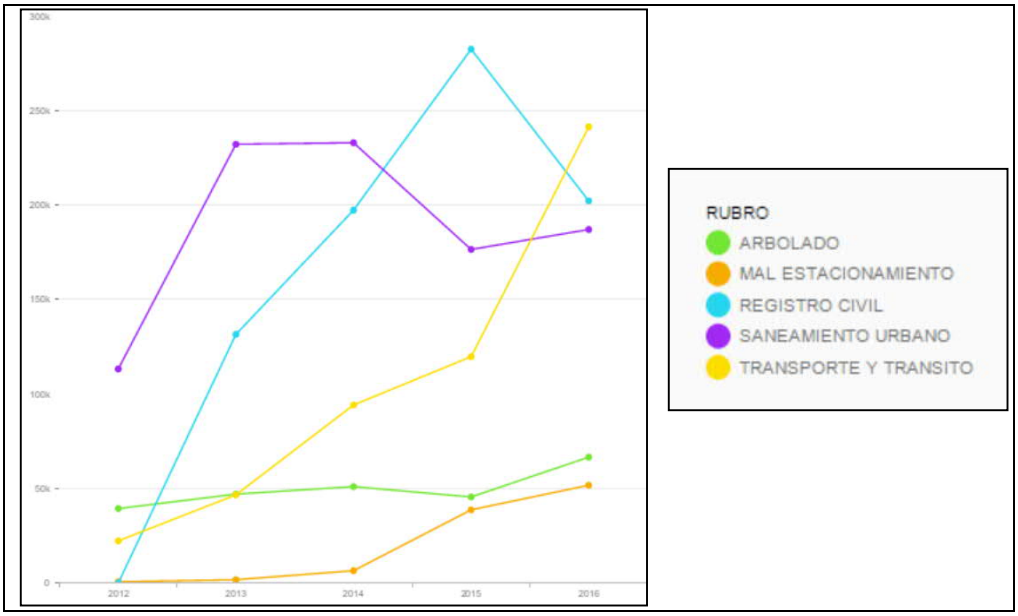
[2] Disponible las 24 hs. en https://gestioncolaborativa.buenosaires.gob.ar/prestaciones

3.1.2 Análisis De La Información

Este análisis surge como resultado del procesamiento de datos sobre una muestra de 3.538.532 solicitudes ingresadas al SUACI, entre los años 2012 y 2016, disponibles en la página del open data del Gobierno de la Ciudad de Buenos Aires https://data.buenosaires.gob.ar.

En el siguiente gráfico podemos apreciar que el rubro Saneamiento Urbano, se mantuvo dentro de los primeros 5 desde el año 2012. Pese a que los rubros Transporte y Tránsito y Registro Civil desplazaron al de Saneamiento Urbano del tope de las solicitudes generadas durante 2016, es una necesidad de la población que persiste año tras año.

Figura 17: Top 5 solicitudes generadas en el SUACI por año.



Fuente: Elaboración propia.

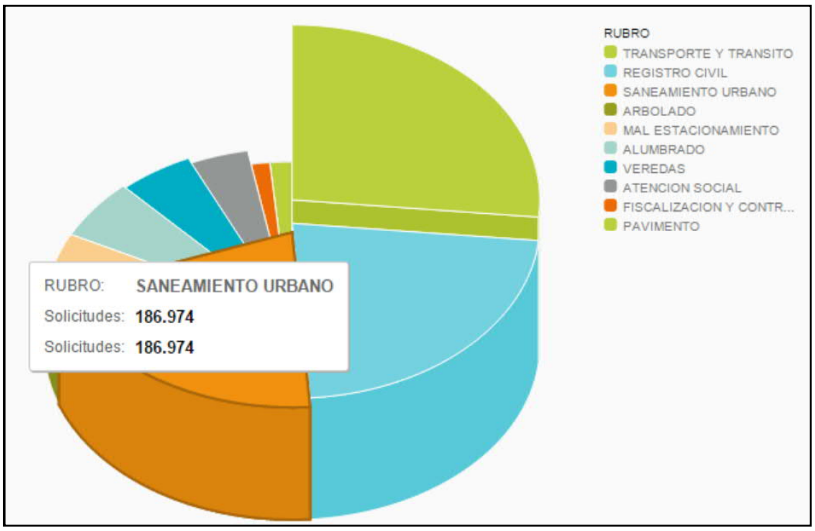
Las solicitudes realizadas en el año 2016 alcanzan un total de 241.286, correspondientes al rubro Transporte y Tránsito, 202.137 para aquellas relacionadas con el rubro Registro Civil y 186.974 para Saneamiento Urbano.

El motivo por el que el rubro Saneamiento Urbano se mantiene dentro de los principales se debe a que existen tipos de residuos especiales que requieren un tratamiento diferencial y que no pueden ser desechados de la manera convencional, esto tipos de residuo pueden ser: escombros, residuos voluminosos y restos de poda, que deben ser retirados a pedido.

Para poder entender un poco más las necesidades de recolección, este análisis se centra en el rubro Saneamiento Urbano, que es el que contiene la información relacionada con el tratamiento de este tipo de residuos.

Como se muestra en la se realizaron un total de 186.974 solicitudes en el año 2016 bajo el rubro Saneamiento Urbano.

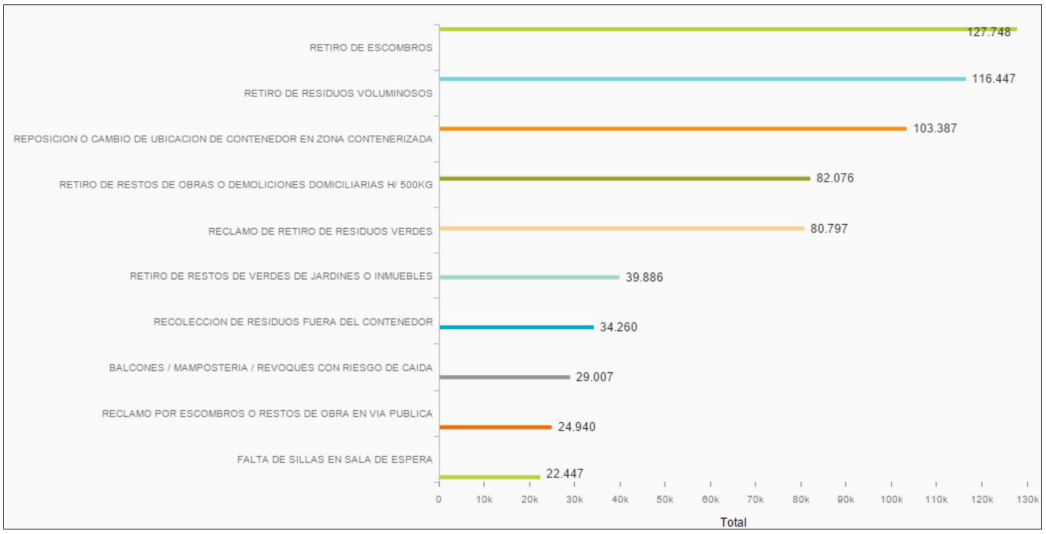
Figura 18: Solicitudes al SUACI por Rubro Año 2016



Fuente: Elaboración propia.

Al analizar el detalle de estas solicitudes, se puede apreciar cuales son los principales motivos de las consultas.

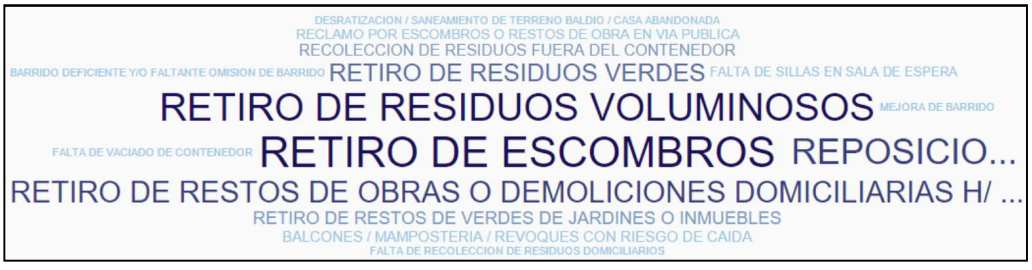
Figura 19: Top 10: Solicitudes al SUACI generadas en el año 2016



Fuente: Elaboración propia.

La mayor cantidad de solicitudes están relacionadas con el retiro de residuos especiales, el siguiente mapa conceptual muestra el foco de las solicitudes generadas durante el año 2016.

Figura 20: Mapa conceptual de Solicitudes al SUACI en el año 2016



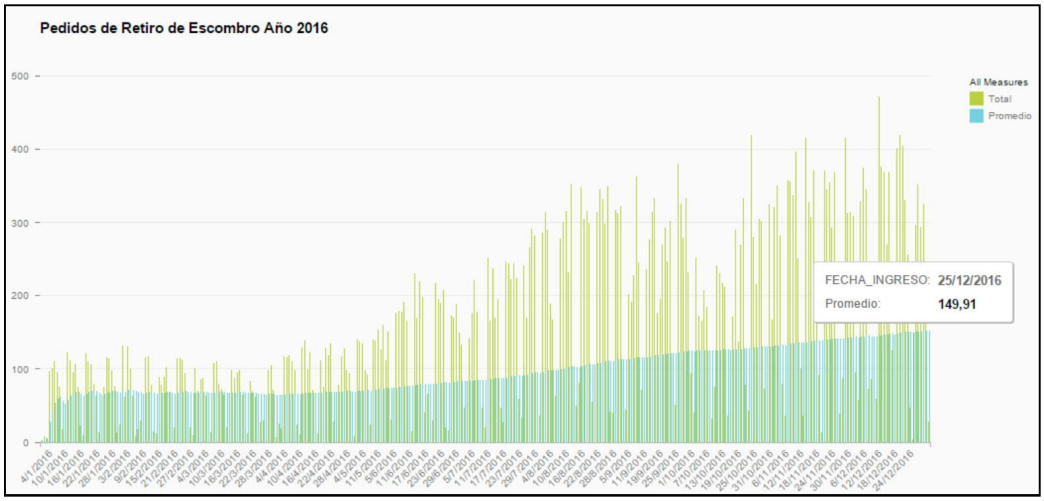
Fuente: Elaboración propia.

3.1.3 Dificultades Encontradas En La Recolección De Residuos

Los residuos especiales (voluminosos, escombros y restos de poda) no deben colocarse en los contenedores, sino que deben de depositarse en la vereda, y se debe contactar al SUACI y para coordinar su retiro de forma gratuita y dentro de las 48 Hs. de realizada la solicitud.

Existe una limitación en la cantidad de escombros que se puede a desechar: Un máximo de 500 Kg, distribuidos en hasta 15 bolsas de escombros comunes, si se excede de esa cantidad se debe contratar un servicio privado. (Gobierno de La Ciudad de Buenos Aires, 2017)

Figura 21: Pedidos al SUACI de Retiro de Escombro por Fecha - Año 2016



Fuente: Elaboración propia.

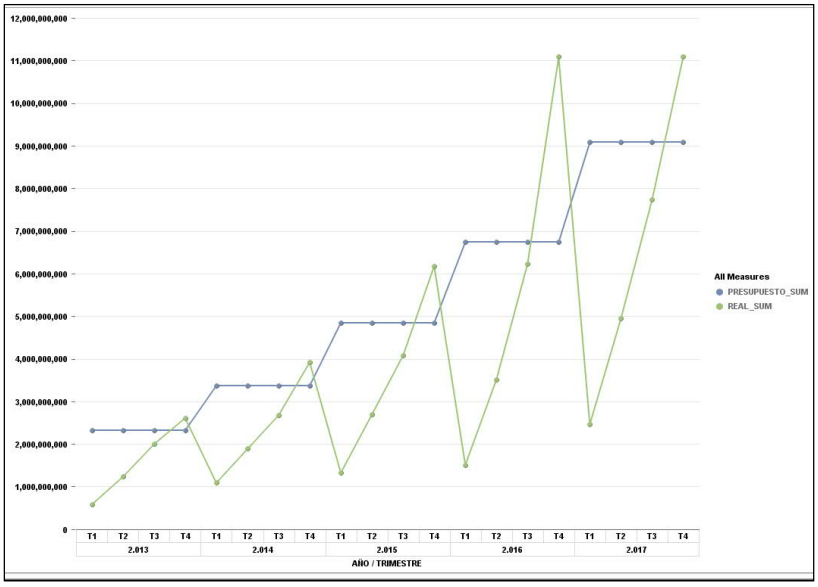
Este gráfico muestra el total de pedidos diarios realizados al SUACI durante el año 2016, en el mismo podemos apreciar un máximo de 472 pedidos realizados el día 12/12/2016, teniendo en cuenta el máximo permitido para desechar, podemos estimar que ese día hubo hasta 236.000 Kg de escombros en las veredas de la Ciudad de Buenos Aires. Considerando el promedio semanal cercano a esa fecha, podemos estimar un

total de 149 pedidos pendientes de recolección, lo que significa un total de hasta 74.500 Kg de escombros que se depositan por día en las veredas de la ciudad. Esta cifra se puede incrementar si el retiro se demora. Pudiendo duplicarse si el retiro se demora hasta el máximo permitido, que es de 48 Hs.

3.1.4 Análisis Del Presupuesto

Para analizar los datos de presupuesto se realiza una comparación entre el presupuesto sancionado y el ejecutado (plan vs real) destinado a la recolección de residuos entre los años 2013 y 2017, la información se encuentra presentada de manera trimestral. Esta comparación se puede visualizar en el siguiente gráfico:

Figura 22: Presupuesto Sancionado Vs Ejecutado (2013-2017)



Fuente: Elaboración propia.

El analisis de este gráfico permite deducir:

1) El presupuesto ejecutado supera al sancionado sistemáticamente durante el último trimestre de cada año. La diferencia es mayor cada año.

2) El presupuesto sancionado es mayor cada año. Esto puede deberse a diversos factores, como por ejemplo la inflación. Lo que hace que la planificación sea mas difícil.
3) Durante el ultimo año del análisis (2017), el presupuesto inicial sancionado (≈ $9.000.000.000) fue menor al total del presupuesto ejecutado del año 2016 (≈11.000.000.000) lo que ya anticipaba que el monto solicitado no iba a ser suficiente para el periodo actual.

3.1.5 Selección De Indicadores

En base a la información analizada y la bibliografía consultada, se proponen los siguientes indicadores para el tablero de control:

Pronóstico De Recolección De Residuos En CABA.

Para analizar el nivel de cumplimiento del objetivo de Basura Cero se propone utilizar la información histórica de la cantidad de residuos recolectados en la CABA.

Para ello se generará una visualización gráfica de datos de recolección de residuos históricos discriminados por comuna, que mediante algoritmos de Machine Learning realicen un pronóstico de valores futuros. De esta forma se brindará a la gerencia información anticipada que le permitirá tomar medidas correctivas proactivamente.

Planificación/Evolución Del Presupuesto

Como indica Kaplan (1999), las medidas financieras no son indicadores relevantes para una empresa del sector público, las mismas funcionan como una limitante y no como un objetivo. Este tipo de organizaciones debe supervisar los gastos, cumplir con el presupuesto financiero asignado y miden su éxito en base a la efectividad con la que administran y la satisfacción de los ciudadanos.

Mediante este indicador se pretende estimar el presupuesto futuro requerido, basándose en información histórica y en el comportamiento del presupuesto consumido.

Esta información permitirá predecir valores futuros, para generar una estimación más precisa, con el objetivo de reducir las diferencias entre el presupuesto planificado y el ejecutado.

Agrupación De Solicitudes Por Concepto Y Comuna

En base al creciente interés y necesidad de la población el tablero de control procesará estas solicitudes, analizando y agrupando las mismas por conceptos similares. Utilizando la referencia geográfica para permitir la gestión de solicitudes afines como un conjunto y detectando fácilmente las zonas de mayor ocurrencia mediante la ubicación en el mapa de la ciudad.

Este indicador funcionará como complemento al indicador financiero, permitirá evaluar las solicitudes existentes y proporcionará una herramientas para simplificar el análisis, con el objetivo de disminuir los tiempos de respuesta y aumentar la satisfacción de la población.

3.2 Diseño del modelo.

3.2.1 Proyección De Recolección De Residuos

Recolección De Datos

Los datos fueron recolectados de la página del Gobierno de la ciudad de Buenos Aires, a través de la iniciativa Buenos Aires Open Data (https://data.buenosaires.gob.ar).

El conjunto de datos seleccionado es el de Recolección de residuos sólidos secos.[3]

Preparación De Datos

El archivo que contiene la información de los residuos recolectados posee la estructura que se detalla a continuación:

Tabla 3: Residuos recolectados por tipo. Ciudad de Buenos Aires.

	Total	Domiciliario	Barrido	Resto	Relleno sanitario
2007	**1.534.803**	**831.202**	**186.554**	**463.034**	**54.012**
Enero	133.070	67.815	17.490	41.122	6.643
Febrero	120.866	61.088	15.766	36.548	7.464
Marzo	141.809	74.069	19.043	43.440	5.257
Abril	134.643	70.314	17.931	42.127	4.272
Mayo	135.719	70.821	18.171	41.278	5.449
Junio	129.017	66.586	15.491	42.937	4.004
Julio	130.541	67.824	14.473	43.963	4.281
Agosto	136.173	67.752	14.836	48.978	4.607
Septiembre	116.256	68.341	14.599	29.027	4.290
Octubre	122.688	74.215	14.044	32.396	2.033
Noviembre	113.710	69.349	12.565	29.736	2.060
Diciembre	120.311	73.029	12.145	31.484	3.653

Fuente: adaptado de Estadísticas y Censos, disponible en http://www.estadisticaciudad.gob.ar/eyc/?p=29141, recuperado el 24 de marzo del 2018.

Es requisito de los algoritmos de Machine Learning que se asigne un identificador (ID) único a cada registro, este ID es la referencia a la serie de tiempo.

Los datos ya se encuentran totalizados por mes y año, pero debe reestructurado para ser analizado adicionando la columna año a cada registro.

Como resultado se obtiene el siguiente formato de archivo, que puede ser importado en el sistema:

[3] Disponible en https://data.buenosaires.gob.ar/dataset/recoleccion-de-residuos-solidos-secos - Consultado el 12/06/2018

Tabla 4: Formato final del archivo de residuos recolectados.

ID	Año	Mes	Total	Domiciliario	Barrido	Resto	Relleno sanitario
1	2007	Enero	133.070	67.815	17.490	41.122	6.643
2	2007	Febrero	120.866	61.088	15.766	36.548	7.464
3	2007	Marzo	141.809	74.069	19.043	43.440	5.257
4	2007	Abril	134.643	70.314	17.931	42.127	4.272
5	2007	Mayo	135.719	70.821	18.171	41.278	5.449
6	2007	Junio	129.017	66.586	15.491	42.937	4.004
7	2007	Julio	130.541	67.824	14.473	43.963	4.281
8	2007	Agosto	136.173	67.752	14.836	48.978	4.607
9	2007	Septiembre	116.256	68.341	14.599	29.027	4.290
10	2007	Octubre	122.688	74.215	14.044	32.396	2.033
11	2007	Noviembre	113.710	69.349	12.565	29.736	2.060
12	2007	Diciembre	120.311	73.029	12.145	31.484	3.653

Fuente: Elaboración propia.

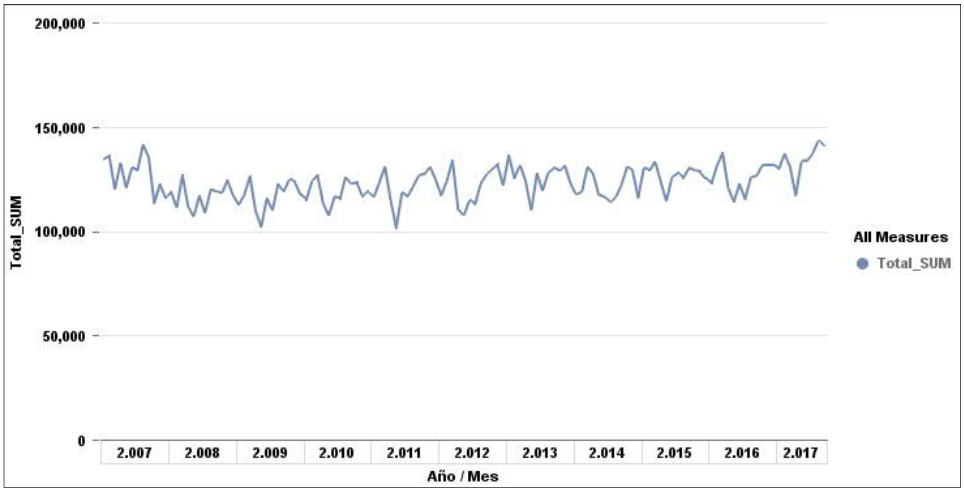
Los datos se importan en la Base de Datos HANA como una tabla columnar, para aprovechar las características de la plataforma.

Como resultado se importaron un total de 129 registros, 120 registros correspondientes a 10 años completos de información 2007 a 2016 y 9 registros correspondientes al año 2017, información disponible en el momento de la importación.

Análisis los datos

Se analizaron los 129 registros que contienen la información de recolección de residuos desde enero 2007 hasta septiembre 2017. Los datos ya se encuentran totalizados por Año/Mes, y se ha agregado el ID correspondiente como referencia de tiempo, por lo que pueden ser procesados por los algoritmos de Machine Learning.

Figura 23: Histórico de Residuos: Total recolectados por año (Tn)



Fuente: Elaboración propia.

Al visualizar la serie de tiempo formada por los datos de recolección de residuos, no se puede detectar a simple vista una tendencia o evolución en la cantidad de residuos generada, por lo que se realiza un análisis de dicha serie de tiempo utilizando las tres variantes de la función Exponential Smoothing.

Entrenamiento Del Modelo: Auto Exponential Smoothing

Para determinar la función más apropiada para el análisis de la serie de tiempo, se realizó un análisis mediante la función Auto Exponential Smoothing, para determinar que función y con qué parámetros se adapta mejor según las características de la serie de tiempo encontrada.

La función Auto Exponential Smoothing, realiza un análisis de la información, probando las diferentes combinaciones de parámetros, para identificar la mejor de acuerdo con los datos proporcionados.

El análisis de los datos por parte del algoritmo da como resultado una estacionalidad de 12 meses en los que se detecta un patrón recurrente en la serie de tiempo lo que sugiere la utilización del algoritmo TESM.

Tabla 5: Resultado función Auto Exponential Smoothing

Parámetro	SESM	DESM	TESM	Descripción
NUMBER_OF_ ITERATIONS	110	160	256	Cantidad de Iteraciones realizadas durante el entrenamiento.
ALPHA	0,234211187	0,567536305	0.6128793897	Parámetros específicos de la Función.
BETA	N/A	0.195077778	0.0000000003	
GAMMA	N/A	N/A	0.0000000003	
CYCLE	N/A	N/A	12	Estacionalidad.
NUMBER_OF_ TRAINING	97	97	97	Cantidad de Registros utilizados para el entrenamiento.
NUMBER_OF_ TESTING	32	32	32	Cantidad de registros utilizados para probar el modelo.

Fuente: Elaboración propia.

Otro análisis realizado sobre la serie de tiempo da como resultado una tendencia positiva. El siguiente gráfico muestra el resultado de este análisis. El parámetro TREND indica los valores -1: Tendencia Negativa, 0: Sin Tendencia, 1: Tendencia Positiva. Esto indicaría que la generación de residuos en la ciudad va en aumento a diferencia de lo esperado por la iniciativa de Basura Cero.

Figura 24: Análisis de Tendencia

	STAT_NAME	STAT_VALUE
1	TREND	1
2	S	2.568
3	P-VALUE	0,00000008652370755793349

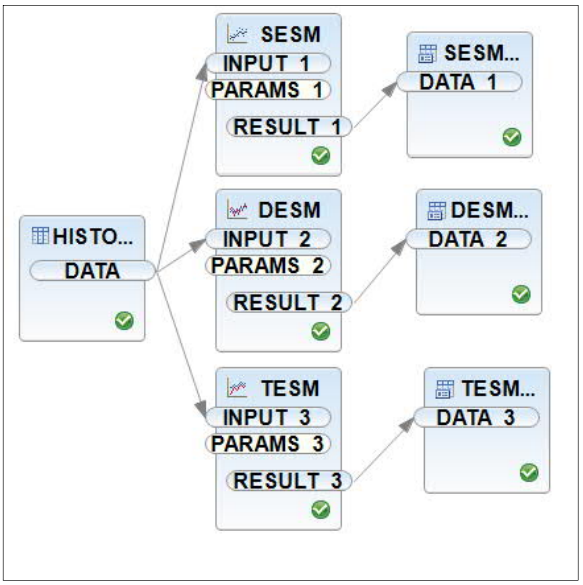
Fuente: Elaboración propia.

Al analizar el resultado de los modelos (SESM, DESM y TESM) se propone como mejor opción al TESM, esto se debe principalmente a la presencia de tendencia y estacionalidad en los datos.

Para entrenar y generar el modelo de datos se genera un flujo de proceso, que es el encargado de orquestar la llamada a las funciones de Machine Learning seleccionadas.

Este flujo es parte de las herramientas provistas por SAP HANA para el procesamiento de la información, utilizando las funciones de PAL.

Figura 25: Flujo de proceso de ML



Fuente: Elaboración propia.

El flujo creado utiliza las 3 funciones predictivas en base a un mismo origen de datos, esta es una funcionalidad que da la plataforma y que permite probar diferentes funciones simplificando la ejecución.

El primer paso consiste en el mapeo de la tabla origen que contiene la información histórica de recolección de residuos, con la estructura que utiliza la función. El campo Domiciliario contiene la información de cantidad de residuos recolectados, expresado en Toneladas (Tn).

La tabla contiene los 129 registros históricos de recolección de residuos en CABA. El mapeo de datos es idéntico para las tres funciones utilizadas.

Figura 26: Mapeo de datos para las funciones SESM/DESM/TESM

Fuente: Elaboración propia.

El segundo paso consiste en la llamada de la función SESM, utilizando los siguientes parámetros:

Parámetros de Entrada:

INPUT: Tabla de datos. Mapeo de datos generado en el Figura 28.

PARAMS: Estructura de parámetros requeridos por la función. La siguiente tabla posee los parámetros utilizados:

Tabla 6: Parámetros de la Función SEMS

Parámetro	Valor	Descripción
ALPHA	0,234211187	Calculado por Auto Exponential Smoothing
FORECAST_NUM	36	Cantidad de valores a predecir (3 Años)
IGNORE_ZERO	1	Ignora los registros en 0

Fuente: Elaboración propia.

Parámetros de Salida:

RESULT: Tabla que contiene el pronóstico generado por la función: SESM_RESULT.

La segunda función ejecutada es DESM, la misma se ejecuta con los siguientes parámetros.

Parámetros de Entrada:

INPUT: Tabla de datos. Mapeo de datos generado en la Figura 28.

PARAMS: Estructura de parámetros utilizados por la función. La siguiente tabla posee los parámetros utilizados:

Tabla 7: Parámetros de la Función DEMS

Parámetro	Valor	Descripción
ALPHA	0,567536305	Calculado por Auto Exponential Smoothing
BETA	0.195077778	Calculado por Auto Exponential Smoothing
FORECAST_NUM	36	Cantidad de valores a predecir (3 años)
IGNORE_ZERO	1	Ignora los registros en 0

Fuente: Elaboración propia.

Parámetros de Salida:

RESULT: Tabla que contiene el pronóstico generado por la función: DESM_RESULT.

La tercera función ejecutada es la función TESM, la función se ejecuta con los siguientes parámetros:

Parámetros de Entrada:

INPUT: Tabla de datos. Mapeo de datos generado en el Figura 28.

PARAMS: Configuración de parámetros para la función. Los parámetros utilizando se detallan en la siguiente tabla.

Tabla 8: Parámetros de la Función TEMS

Parámetro	Valor	Descripción
ALPHA	0.6128793897	Calculado por Auto Exponential Smoothing
BETA	0.0000000003	Calculado por Auto Exponential Smoothing
GAMMA	0.0000000003	Calculado por Auto Exponential Smoothing
CYCLE	12	Calculado por Auto Exponential Smoothing
FORECAST_NUM	36	Cantidad de valores a predecir (3 Años)
IGNORE_ZERO	1	Ignora los registros en 0

Fuente: Elaboración propia.

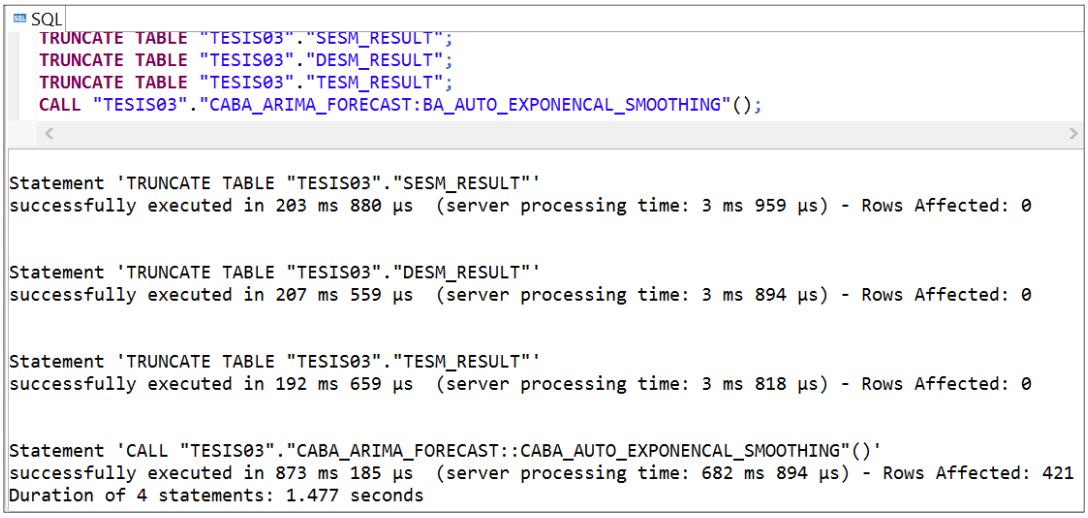
Parámetros de Salida:

RESULT: Tabla que contiene el pronóstico generado por la función: TESM_RESULT.

Prueba Del Modelo

Al activar el modelo se genera un Procedimiento Almacenado (Stored Procedure) de la BD HANA que se puede ejecutar desde una consola de SQL.

Figura 27: Log de ejecución algoritmos SESM/DESM/TESM

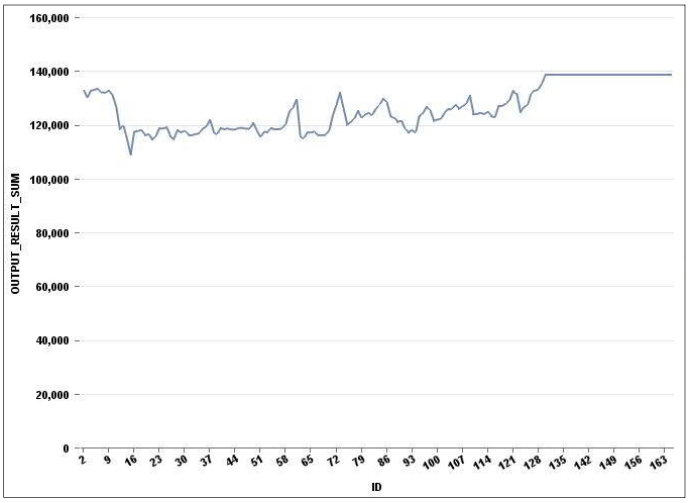


Fuente: Elaboración propia.

La Figura 28 muestra el log de ejecución. En este caso se ejecutaron los tres algoritmos para los 129 registros en 682 ms.

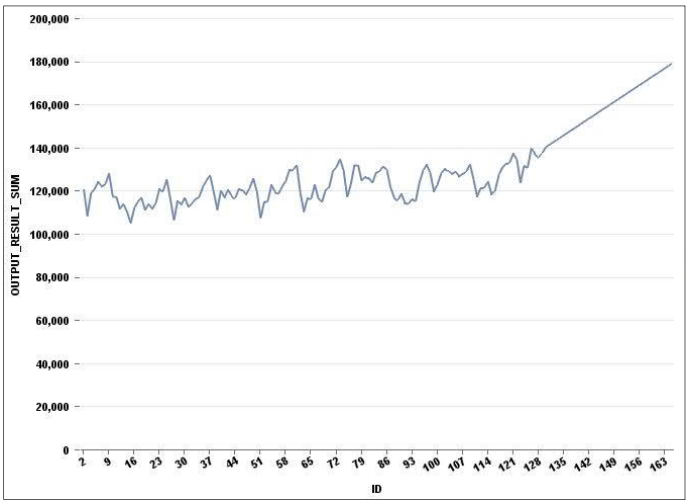
El resultado del pronóstico de las funciones SESM, DESM y TESM se representa de forma gráfica con el objetivo de comparar el resultado de cada uno de los pronósticos.

Figura 28: Resultado función SESM



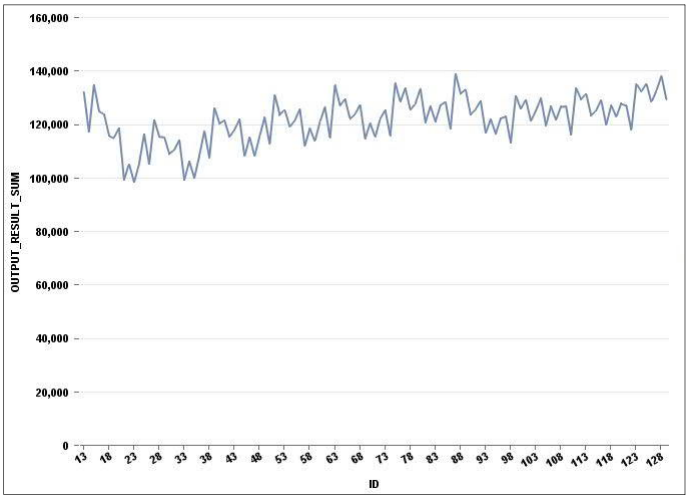
Fuente: Elaboración propia.

Figura 29: Resultado función DESM



Fuente: Elaboración propia.

Figura 30: Resultado función TESM

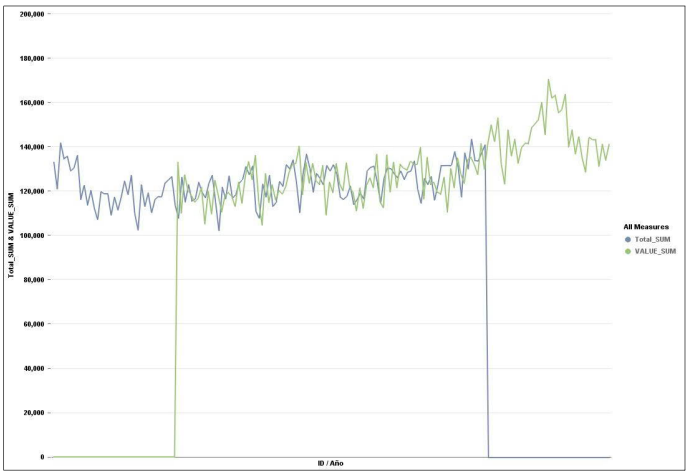


Fuente: Elaboración propia.

Uso Del Modelo

El pronóstico obtenido por las funciones se puede integrar con la información histórica, de esta forma se puede visualizar la evolución esperada. El siguiente gráfico muestra el resultado de unir los datos históricos con el pronóstico generado por la función TESM.

Figura 31: Pronóstico TESM



Fuente: Elaboración propia.

Se puede apreciar en el siguiente gráfico que la predicción (línea verde) respeta la forma de los valores históricos (línea azul), gracias a la incorporación de la estacionalidad y la tendencia detectadas.

3.2.2 Proyección De Presupuesto

Recolección De Datos

Para este escenario se utilizarán el conjunto de datos recolectados desde el repositorio de Buenos Aires Open Data: Presupuesto Ejecutado.[4]

Preparación De Los Datos:

La información sobre el presupuesto ejecutado se encuentra almacenada en varios archivos. Los archivos se denominan Presupuesto-Ejecutado-[año].csv, donde [año] indica el año al que pertenece la información.

El archivo posee una en una estructura de árbol, y contiene la información de las diferentes obras que fueron ejecutándose en el año. Para el análisis se toma solamente aquella información de Higiene Urbana en el subgrupo Recolección Y Limpieza Por Terceros, donde se encuentra la información discriminada por Comuna.

La información se importa a la BD HANA en una tabla columnar con los siguientes ajustes:

- Se generó un ID secuencial para cada registro.
- Se suprimieron los campos no relevantes (para simplificar el análisis)

Luego se genera una vista de BD para totalizar la información por año y trimestre:

---

[4] Obtenido de https://data.buenosaires.gob.ar/dataset/presupuesto-ejecutado. Consultado el 11/06/2018

Figura 32: Creación de vista de presupuesto ejecutado

```sql
/* Crea una secuencia para asignar el ID */
DROP SEQUENCE ARIMA_ID;
CREATE SEQUENCE ARIMA_ID MINVALUE 0;

-- CREA UNA VISTA REDUCIDA DE LOS DATOS DE PRESUPUESTO
DROP TABLE "TESIS03"."PRESUPUESTO_EJECUTADO_ARIMA";
CREATE COLUMN TABLE "TESIS03"."PRESUPUESTO_EJECUTADO_ARIMA" AS
(
select  ARIMA_ID.NEXTVAL AS "ID",          --GENERA EL ID PARA CADA REGISTRO
        "AÑO", "TRIMESTRE",                --CLAVE ANTERIOR
        sum("SANCION")    AS SUM_SANCION, -- TOTALES CALCULADOS
        sum("VIGENTE")    AS SUM_VIGENTE,
        sum("DEFINITIVO") AS SUM_DEFINITIVO,
        sum("DEVENGADO")  AS SUM_DEVENGADO
  from "TESIS03"."PRESUPUESTO_EJECUTADO"
 group by "AÑO", "TRIMESTRE"
)
```

Fuente: Elaboración propia.

Analisis De Los Datos

Entrenamiento Del Modelo: ARIMA

Para realizar el pronóstico utilizando las funciones ARIMA, se dispone de dos funciones: AUTOARIMA y ARIMA_FORECAST.

AUTOARIMA:

Aunque el modelo ARIMA es útil y poderoso en el análisis de series de tiempo, de alguna manera es difícil elegir el orden apropiado de los parámetros. La función AUTOARIMA identifica los parámetros de un modelo ARIMA automáticamente y genera el modelo de datos que sirve como entrada para la función ARIMA_FORECAST.

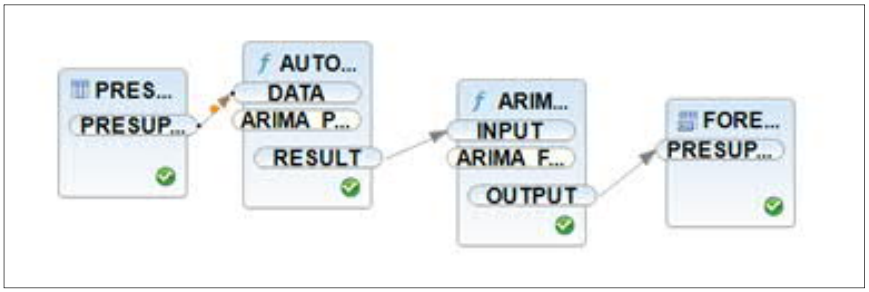
ARIMA_FORECAST

Esta función crea la serie de tiempo en base al modelo generado por la función AUTOARIMA.

Flujo De Proceso

El flujo posee 4 pasos y realiza la llamada a 2 funciones de PAL. Los pasos se detallan a continuación:

Figura 33: Flujo de entrenamiento Función AUTOARIMA

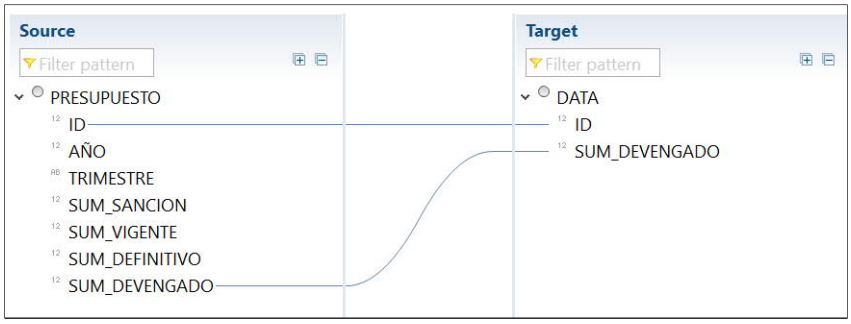


Fuente: Elaboración propia.

El primer paso consiste en el mapeo de la tabla que contiene los datos de presupuesto ejecutado, utilizados como datos de entrenamiento, con la estructura que requiere la función AUTOARIMA.

El mapeo se realiza mediante una funcionalidad gráfica como se muestra en la Figura 33.

Figura 34: Mapeo de datos función AUTOARIMA



Fuente: Elaboración propia.

El segundo paso consiste en la llamada a la función AUTOARIMA provista por la plataforma. La función contiene un total de 3 parámetros; dos parámetros de entrada y uno de salida:

Parámetros de Entrada:

DATA: Datos a ser analizados: Mapeados en la Figura 31.

ARIMA_PARAMS: Estructura de parámetros que permite definir el comportamiento de la función.

Tabla 9: Parámetros de entrada AUTOARIMA

Parámetro	Valor	Descripción
SEASONAL_PERIOD	-1	Valor del periodo estacional. El valor -1 indica que se calcula automáticamente.
SEASONALITY_CRITERION	0,1	El criterio del coeficiente de autocorrelación para aceptar la estacionalidad, en el rango de (0, 1). Cuanto más grande es, menos probable es que una serie temporal se considere estacional.
D	-1	Orden de primera diferenciación. Negativo: Identifica automáticamente el primer orden de diferenciación con la prueba KPSS. Otros: Utiliza el valor especificado como primer orden de diferenciación.
KPSS_SIGNIFICANCE_LEVEL	0,05	El nivel de significación para la prueba KPSS. Los valores admitidos son 0,01, 0,025, 0,05 y 0,1. Cuanto más pequeña es, mayor es la probabilidad de que una serie temporal se considere como primera estacionaria, es decir, menos probable que necesite una primera diferenciación.
MAX_D	2	El valor máximo de D cuando se aplica la prueba KPSS.

Parámetro	Valor	Descripción
SEASONAL_D	-1	Orden de las diferencias estacionales. Negativo: Identifica automáticamente el orden de diferenciación estacional de la prueba Canova-Hansen. Otros: Utiliza el valor especificado como orden de diferenciación estacional.
CH_SIGNIFICANCE_LEVEL	0,05	El nivel de significación para la prueba Canova-Hansen. Los valores admitidos son 0,01, 0,025, 0,05, 0,1 y 0,2. Cuanto más pequeña es, mayor es la probabilidad de que una serie temporal se considere estacionaria/estacional, es decir, menos probable que necesite una diferenciación estacional.
MAX_SEASONAL_D	2	El valor máximo de SEASONAL_D cuando se aplica la prueba Canova-Hansen.
MAX_P	2	El valor máximo de p para la orden AR.
MAX_Q	2	El valor máximo de q para la orden MA.
MAX_SEASONAL_P	2	El valor máximo de p para la orden AR. (estacionario)
MAX_SEASONAL_Q	2	El valor máximo de q para la orden MA.(estacionario)

Fuente: Elaboración propia.

Los valores utilizados fueron ajustados en base al resultado de las pruebas y las recomendaciones dadas por el algoritmo AUTO_ARIMA. El proceso realizado consiste en tomar inicialmente valores amplios, para dejar que el algoritmo calcule los más

óptimos para la serie de tiempo analizada, al ajustar los parámetros se consigue la mejor estimación en menor tiempo.

Parámetros de Salida:

RESULT: retorna el modelo generado para los datos de entrenamiento utilizados.

Figura 35: Modelo Generado AUTOARIMA

NAME	VALUE
p	0
AR	
d	1
q	1
MA	-0.781542
s	4
P	1
SAR	0.892648
D	0
Q	0
SMA	
sigma^2	1.46293e+18
log-likelihood	-427.57
AIC	861.14
AICc	862.74
BIC	863.973
dy(n-p:n-1)_aux	
dy_aux	0
dy_0	6.50701e+08;7.66983e+08;
y(n-d:n-1)_aux	0
y(n-d:n-1)_0	1.10878e+10
epsilon(n-q:n-1)_aux	0
epsilon(n-q:n-1)_0	-8.711e+08

Fuente: Elaboración propia.

Se utiliza la función ARIMA_FORECAST con el modelo generado por la función AUTOARIMA, para pronosticar la evolución del pronóstico durante los próximos 3 años (36 interacciones). La función ARIMA_FORECAST posee 3 parámetros; 2 de entrada y 1 de salida:

Parámetros de Entrada:

INPUT: Modelo generado en el paso por la función AUTOARIMA.

PARAMS: Estructura de parámetros para la función.

Mediante esta estructura se indica el parámetro ForecastLength = 36 para pronosticar los próximos 3 años.

Parámetros de Salida:

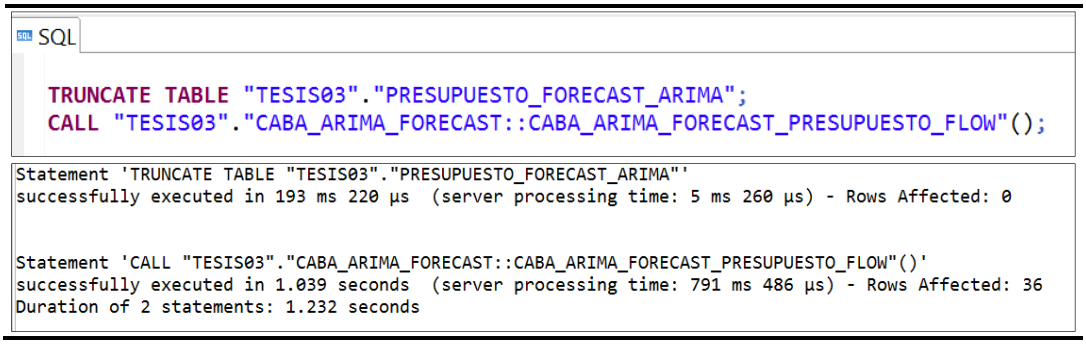
RESULT: Pronóstico generado por la función. Esta información es almacenada en una nueva tabla.

El resultado se almacena en una tabla columnar para su posterior análisis.

Prueba Del Modelo: ARIMA

La siguiente figura muestra secuencia de comandos realizados para llamar al procedimiento almacenado que contiene el flujo de proceso generado. El tiempo de procesamiento en el servidor es de 791 ms. Para generar el modelo y realizar el pronóstico de 36 meses.

Figura 36: Secuencia de comandos SQL y Log de Ejecución



Fuente: Elaboración propia.

Una vez generado el modelo se grafica utilizando la funcionalidad de Análisis de datos de HANA sobre la tabla de datos pronosticados.

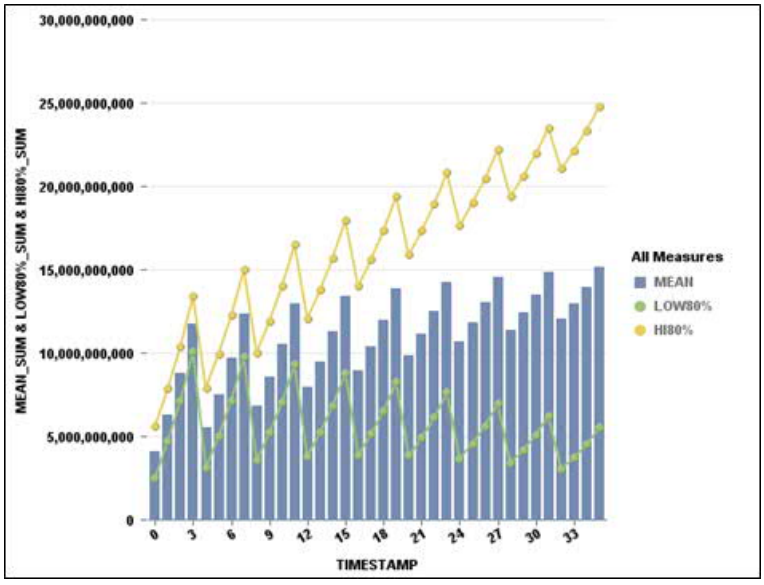
La funcion ARIMA_FORECAST pronostíca el resultado utilizando la siguiente estructura

MEAN: Valor estimado por la función.

LOW80%: Determina el valor mínimo del intervalo donde se estima con un 80% de exactitud que el valor pronósticado se encontrará.

HIGH80%: Determina el valor máximo del intervalo donde se estima con un 80% de exactitud que el valor pronósticado se encontrará.

Figura 37: Resultado de la Función ARIMA



Fuente: Elaboración propia.

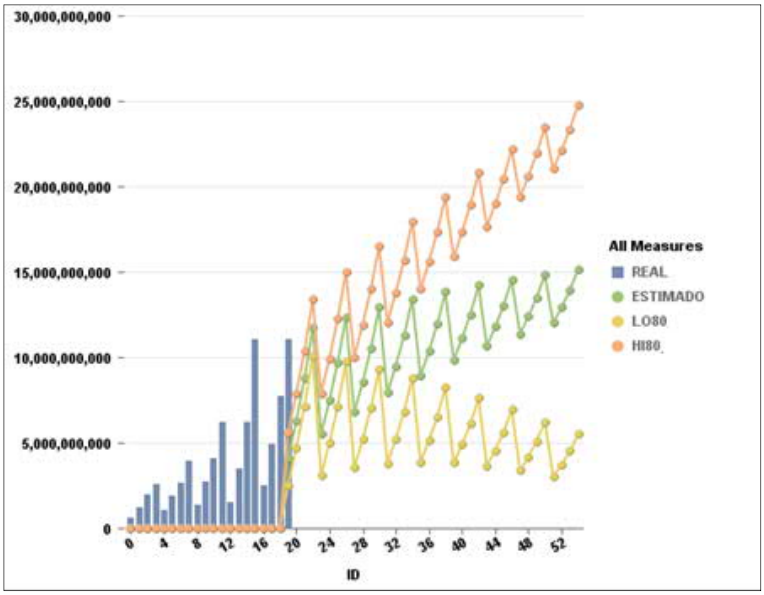
LOW95%: Determina el valor mínimo del intervalo donde se estima con un 95% de exactitud que el valor pronósticado se encontrará.

HIGH95%: Determina el valor máximo del intervalo donde se estima con un 95% de exactitud que el valor pronósticado se encontrará

Uso Del Modelo

Finalmente se analiza la información se contrasta con los valores históricos.

Figura 38: Proyección de presupuesto ARIMA



Fuente: Elaboración propia.

Mediante el uso de los valores LOW80/HIGH80 se permite generar un rango de valores en los cuales se estima que existe el 80% de posibilidad de que el valor real se encuentre dentro de ese rango.

3.2.3 Agrupación de Solicitudes

Recolección De Datos

Los datos fueron proporcionados desde el Gobierno de la ciudad de Buenos Aires, a través de la iniciativa Buenos Aires Open Data (https://data.buenosaires.gob.ar):

Para este análisis se utilizan los Contactos al Sistema Único de Atención Ciudadana correspondientes al año 2017, disponibles en abril 2018.

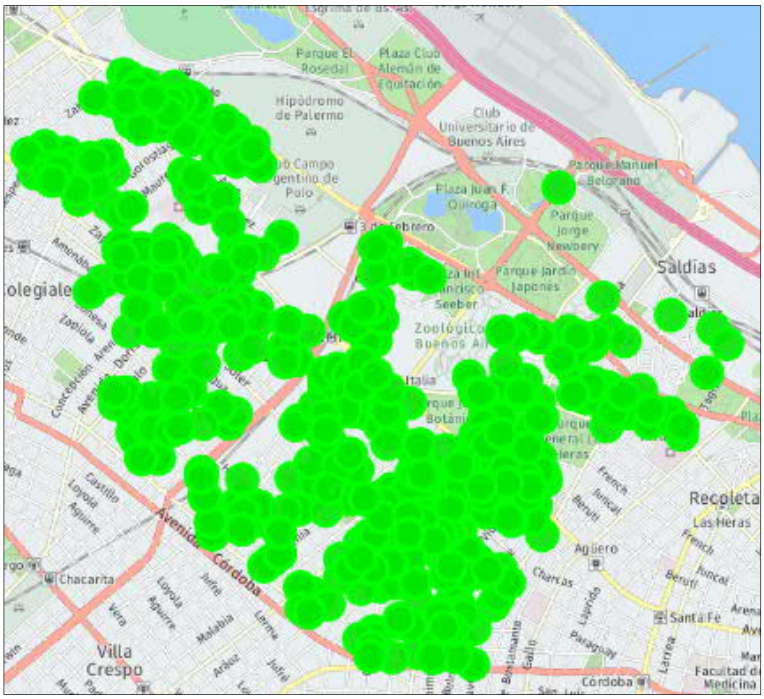
Preparación De Datos

Para permitir al motor espacial analizar la información geográfica, se requiere unificar los campos Latitud y Longitud que contiene cada registro, para formar un punto geográfico almacenado dentro de un campo de tipo ST_Point. Esto permite realizar las operaciones espaciales, así como también utilizar los algoritmos de agrupación.

Análisis De Datos

El atributo principal de esta información es la ubicación geográfica de cada una de las solicitudes, este atributo, en conjunto con el concepto y la comuna a la que pertenece, permite agrupar las solicitudes por afinidad. El siguiente gráfico muestra la distribución geográfica de una muestra de solicitudes al SUACI.

Figura 39: Distribución geográfica de las solicitudes



Fuente: Elaboración propia.

Entrenamiento De Datos

En este conjunto de datos se utiliza un algoritmo no supervisado, por lo que no es necesario el paso de entrenamiento.

Prueba Del Modelo

Se procesa el conjunto de datos utilizando el algoritmo K-Means definiendo un total de 10 grupos para un total de 253.662 registros.

El proceso de las solicitudes genera los grupos solicitados con su correspondiente centroide, expresado como un punto geográfico, y la cantidad de solicitudes que contiene ese grupo, la siguiente figura muestra el resultado del procesamiento:

Figura 40: Resultado de la agrupación de solicitudes

	CENTROIDCLUSTER	COUNTER
1	{"type": "Point", "coordinates": [-34.58566, -58.43984]}	22.911
2	{"type": "Point", "coordinates": [-34.595732, -58.4082775]}	28.823
3	{"type": "Point", "coordinates": [-34.558276, -58.461279]}	24.746
4	{"type": "Point", "coordinates": [-34.630598, -58.425974]}	27.520
5	{"type": "Point", "coordinates": [-34.621767, -58.46145]}	32.121
6	{"type": "Point", "coordinates": [-34.635901, -58.508397]}	28.555
7	{"type": "Point", "coordinates": [-34.620784, -58.382972]}	21.353
8	{"type": "Point", "coordinates": [-34.670181, -58.478052]}	15.356
9	{"type": "Point", "coordinates": [-34.603124, -58.501009]}	29.998
10	{"type": "Point", "coordinates": [-34.572745, -58.485806]}	22.279

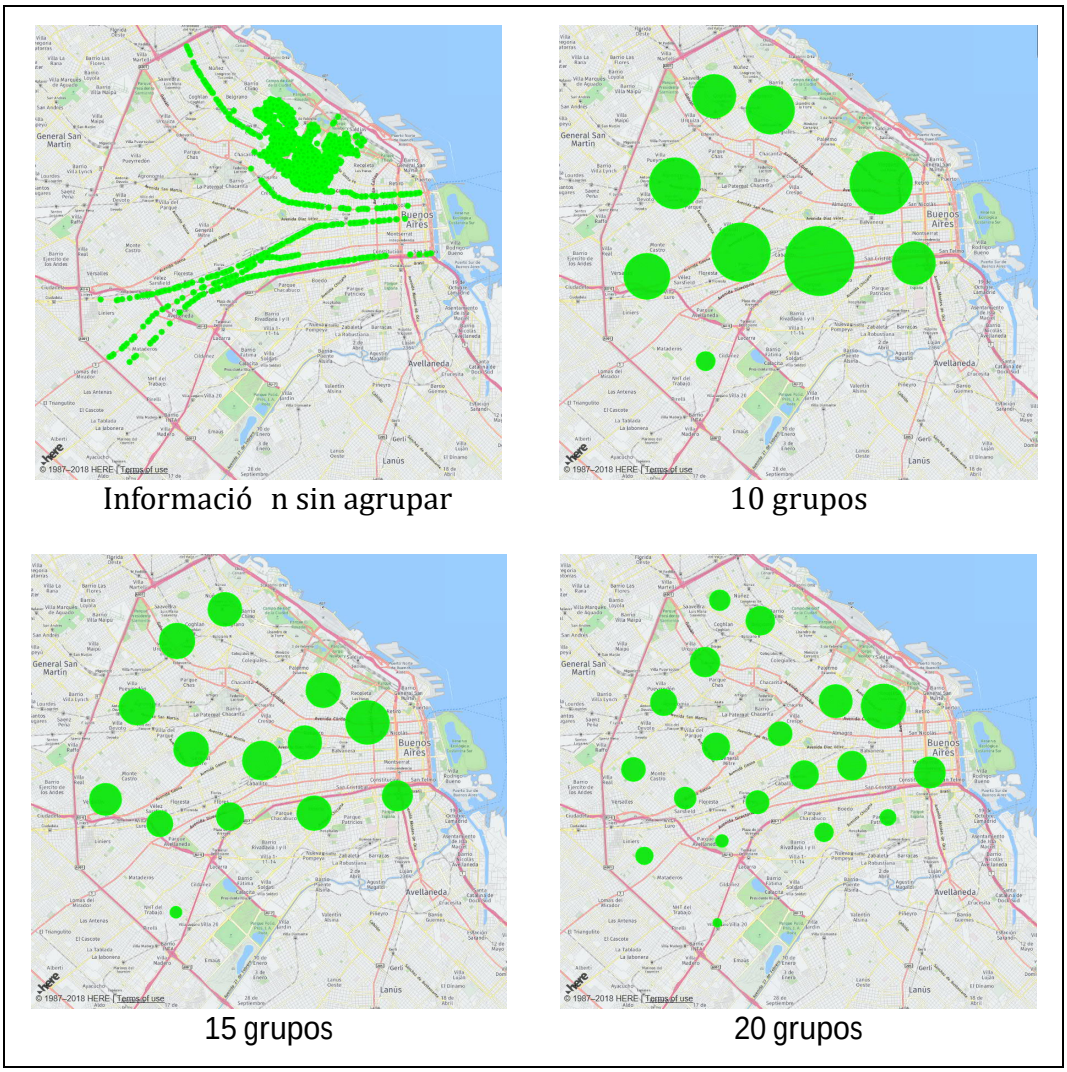
Fuente: Elaboración propia.

Uso Del Modelo

La información se utiliza para graficar en el mapa de la Ciudad Autónoma de Buenos Aries los grupos generados en base a los diferentes conceptos de cada solicitud. El siguiente ejemplo muestra las solicitudes correspondientes al concepto RETIRO DE

ESCOMBRO / RESTO DE OBRA utilizando valores diferentes de K para formar los grupos. Cada círculo dibujado representa un grupo definido por el algoritmo K-Means.

Figura 41: Agrupación de solicitudes utilizando K-Means



Fuente: Elaboración propia

3.2.4 Otros Datos

Con el objetivo de proveer más información sobre las comunas se incluyeron datos relacionados con las mismas, en base a la información proporcionada por Estadísticas y Censos. Esta información complementa los datos del tablero de gestión.

Origen: Proyecciones de población por comuna y sexo. Ciudad de Buenos Aires. Años 2010/2025.

Fuente http://www.estadisticaciudad.gob.ar/eyc/?p=28146 accedido el 17/08/2018.

Las características de los datos son:

- **Área Temática**: Población
- **Tema:** Estructura de la Población
- **SubTema:** Sexo y Edad
- **Año Desde:** 2010
- **Año Hasta:** 2025
- **Distribución geográfica:** Comuna
- **Fuente:** Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA). Proyecciones de población.

Esta información se utiliza para conocer la cantidad de habitantes por caca Comuna.

Origen: Residuos recolectados por tipo y promedio diario por habitante. Ciudad de Buenos Aires. Años 1995/2016

Fuente: http://www.estadisticaciudad.gob.ar/eyc/?p=29140 Accedido el 6/11/2017.

Las características de los datos son:

- **Área Temática:** Servicios Públicos
- **Tema:** Residuos
- **Año Desde:** 1995
- **Año Hasta:** 2016

- **Distribución geográfica:** Total Ciudad
- **Fuente:** Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA) sobre base de datos de CEAMSE.

Esta información es utilizada en el prototipo para detallar la cantidad de residuos generados por habitante.

Origen: Comunas de la Ciudad incluyendo los barrios que la componen.

Fuente: https://data.buenosaires.gob.ar/dataset/comunas - Accedido el 15/01/2018

Las características de los datos son:

- **Área Temática:** Información Geográfica
- **Tema:** Comunas Ciudad
- **Fuente del dato:** Publicado por Ministerio de Modernización, Innovación y Tecnología - SS de Innovación y Ciudad Inteligente - DG de Gestión Digital Unidad de Sistemas de Información Geográfica (USIG)
- **Fecha de relevamiento de los datos:** 31/12/2015
- **Fecha de Actualización del dato:** abril 2008

Esta información contiene la información geográfica de cada comuna, expresada como un polígono múltiple, que representa los límites de cada comuna.

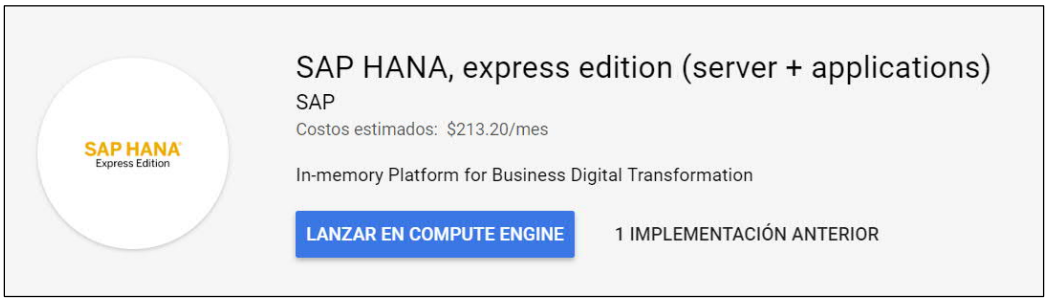
3.3 Arquitectura De La Solución

La solución propuesta está a basada en Google Cloud Platform (GCP) que es la plataforma de servicios Cloud que ofrece Google LLC[5].

GCP provee una imagen predefinida de SAP HANA Express Edition disponible en sus dos opciones de instalación:

- Solo servidor: incluye solo el servidor de BD.
- Servidor + aplicaciones: incluye el servidor de BD + motor XS (Opción utilizada para el prototipo)

Figura 42: GCP Imagen disponible SAP HANA Express Edition



Fuente: de SAP HANA, express edition, in Google Cloud Platform Launcher (Database Services), disponible en https://www.sap.com/developer/tutorials/hxe-gcp-getting-started-launcher.html, recuperado el 17 de agosto de 2018.

La versión de HANA Express Edition es libre y gratuita, con una limitación de 32 Gb de Memoria. GCP trabaja con el concepto de BYOL (Bring Your Own License) por lo que esta implementación no posee costos adicionales por el software. La siguiente figura muestra un estimado del costo del servidor en un periodo de 30 días con un uso de 24 horas.

[5] © 2017 Google LLC Todos los derechos reservados. Google y el logotipo de Google son marcas comerciales registradas de Google LLC

Figura 43: Costos GCP mensual estimados

Elemento	Costos estimados
Tarifa por uso de SUSE Linux Enterprise 12 Mostrar más	$80.30/mes
Costos de Google Compute Engine	
Instancia de VM: 4 CPU virtuales + 26 GB de memoria (n1-highmem-4)	$172.86/mes
Disco de estado sólido: 70 GB	$11.90/mes
Descuento por uso continuo	- $51.86/mes
Total	$213.20/mes

Fuente: de Precios de Google Compute Engine, disponible en https://cloud.google.com/compute/pricing?hl=es_419&_ga=2.40093728.-34433375.1525622736 , recuperado el 17 de agosto 2018 (requiere Login)

3.3.1 ¿Por Qué Elegir Google Cloud Platform?

Google lleva más de 15 años desarrollando su infraestructura en la Nube. Google usa esta infraestructura internamente en varios servicios altamente utilizados a escala global, entre ellos Gmail, Google Maps, YouTube y el motor de búsqueda. Debido al tamaño y la escala de estos servicios, Google destinó mucho trabajo a optimizar su infraestructura y crear un conjunto de herramientas y servicios para administrarla de forma efectiva. Google Cloud Platform pone esta infraestructura y estos recursos de administración disponibles para el uso en general.

GCP Provee un periodo de prueba de 12 meses, durante este periodo se otorga un monto de $300 dólares, a utilizar cualquier producto de GCP. Esta opción está disponible únicamente para nuevos clientes y por única vez.[6]

[6] https://cloud.google.com/free/

La siguiente muestra los diferentes componentes que forman parte de la solución.

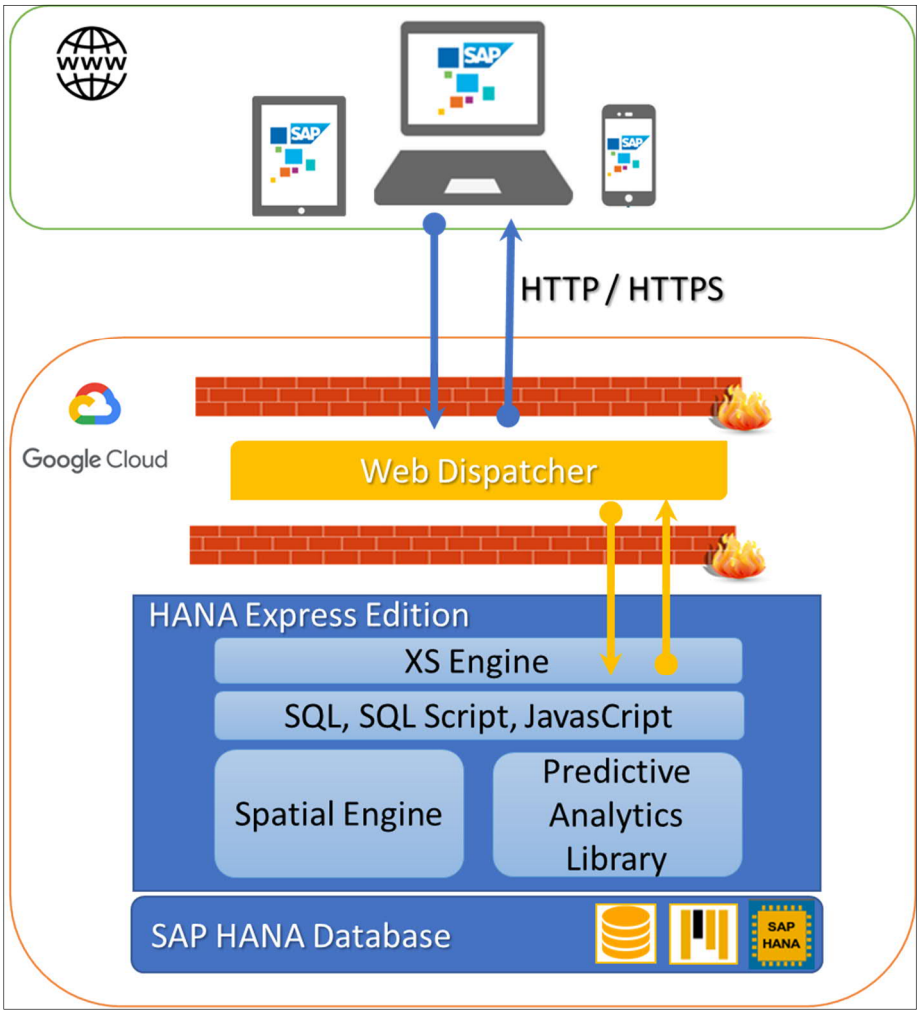
3.3.2 Vista Física

Los componentes de arquitectura de la solución propuesta son:

Acceso Usuario Final

El acceso del usuario final se realiza desde Internet utilizando protocolo HTTP/HTTPS accedido desde un explorador de Internet.

Figura 44: Arquitectura de la solución



Fuente: Elaboración propia.

Cortafuegos (Firewall)

El cortafuegos es un elemento de seguridad diseñado para bloquear los accesos no autorizados, mediante el uso de un cortafuegos se genera una zona desmilitarizada para permitir el acceso a los usuarios desde Internet.

SAP Web Dispatcher[7]

Este producto se ubica entre los clientes Web y los productos SAP que deben ser accedidos desde Internet. Es el punto de entrada de las peticiones HTTP/HTTPs y posee la función de aceptar o denegar las peticiones, funciona también como balanceador de carga, filtro de direcciones URL y redireccionamiento. También puede ser configurado para funcionar como un proxy reverso.

Plataforma HANA Express Edition

XS Engine[8]

El concepto central de SAP HANA Extended Application Services es integrar un servidor de aplicaciones completo, un servidor web y un entorno de desarrollo en la plataforma SAP HANA. No se trata de otro software más instalado en el mismo hardware que SAP HANA, sino que SAP ha decidido integrar esta nueva funcionalidad de servicios de aplicaciones directamente en el núcleo de la base de datos de SAP HANA, ofreciéndole una oportunidad de rendimiento y acceso a las características diferenciadoras de SAP HANA que ningún otro servidor de aplicaciones tiene.

---

[7] SAP Web Dispatcher:
https://help.sap.com/viewer/683d6a1797a34730a6e005d1e8de6f22/7.5.7/en-US

[8] SAP HANA Extended Application Services.
https://blogs.sap.com/2012/11/29/sap-hana-extended-application-services/

SQL, SLQScript, JavaScript[9]

Parte de la plataforma SAP HANA posee un motor de procesamiento de SQL, SQLSctipt, que es una versión adaptada del SQL para SAP HANA y JavaScript.

Spatial Engine[10]

SAP HANA incluye un motor espacial multicapa que admite columnas espaciales, métodos de acceso espacial y sistemas de referencia espacial.

Predictive Analytics Library[11]

La Biblioteca de Análisis Predictivo (PAL) define funciones que pueden ser llamadas desde los procedimientos SQLScript para realizar algoritmos analíticos. Esta versión de PAL incluye algoritmos de análisis predictivo.

SAP HANA Database

Es la base de datos Columnar y en Memoria implementada por SAP SE.

3.3.3 Escalabilidad

La versión SAP HANA Express Edition, no está preparada para escalar, ya que la misma se trata de una versión que no requiere licencias y que está diseñada principalmente para el armado de prototipos, aunque existen también casos de implementaciones productivas de aplicaciones no críticas. Sólo permite extender la capacidad de memoria hasta 128Gb mediante el pago de licencia.

---

[9] SAP HANA Extended Application Services:
https://blogs.sap.com/2012/11/29/sap-hana-extended-application-services/
[10] SAP Spatial Reference:
https://help.sap.com/viewer/cbbbfc20871e4559abfd45a78ad58c02/2.0.02/en-US/4a55e68327a54e9a9c105145186c346f.html
[11] SAP HANA PAL
https://help.sap.com/viewer/2cfbc5cf2bc14f028cfbe2a2bba60a50/2.0.03/en-US

Si es necesario utilizar una memoria mayor a 128 GB o se requiere un entorno de alta disponibilidad, se debe optar por las versiones Empresariales: SAP HANA Platform Edition y SAP HANA Enterprise Edition.

3.3.4 Alta Disponibilidad (HA) Y Recuperación Ante Desastres (DR)

Al igual que con la escalabilidad, la versión de SAP HANA Express Edition, carece las características de alta disponibilidad, a diferencia de las versiones Enterprise, que si provee 3 opciones diferentes. No todas las capacidades de Alta Disponibilidad de SAP HANA están disponibles en Google Cloud Platform (GCP)[12], las opciones se detallan a continuación:

Conmutación Automática De Host

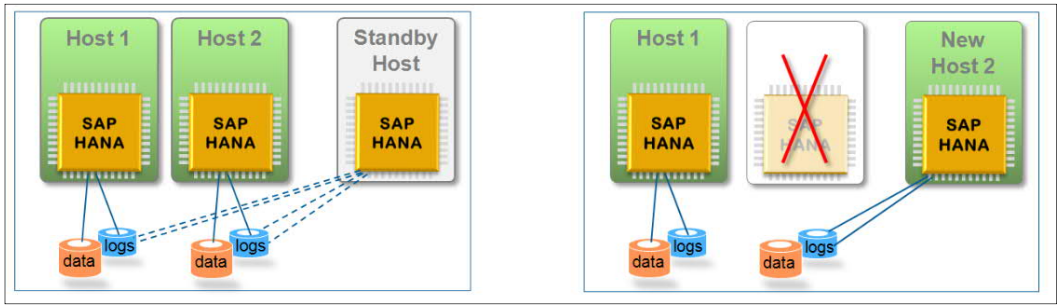
SAP HANA host auto-failover es una solución local de recuperación de fallos que se puede utilizar como medida adicional o alternativa a la replicación del sistema. Se agregan uno (o más) hosts de reserva a un sistema SAP HANA y se configuran para que funcionen en modo de espera. Cuando un host primario falla, un host secundario automáticamente toma su lugar.

SAP HANA host auto-failover no es compatible con GCP. La migración en vivo del motor de computación sirve para el mismo propósito en un sistema SAP HANA en GCP.

---

[12] SAP HANA High Availability and Disaster Recovery Planning Guide
https://cloud.google.com/solutions/partners/sap/sap-hana-hadr-planning-guide

Figura 45: Conmutación automática de Host

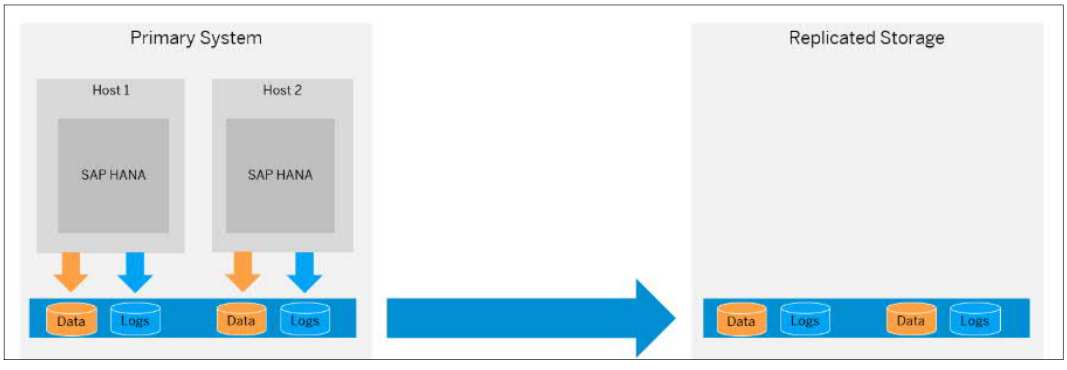


Fuente: SAP HANA HA/DR, disponible en https://www.sap.com/documents/2016/06/f6b3861d-767c-0010-82c7-eda71af511fa.html , recuperado el 15 de agosto de 2018.

Replicación Del Almacenamiento

Una desventaja de las copias de seguridad es la pérdida potencial de datos entre el momento de la última copia de seguridad y el momento de la falla. Por lo tanto, una solución es proporcionar una replicación continua de todos los datos persistentes. La replicación de almacenamiento síncrono sólo se puede utilizar cuando la distancia entre el sitio primario y el de respaldo es relativamente corta (normalmente 100 kilómetros o menos), lo que permite latencias de ida y vuelta de menos de milisegundos.

Figura 46: Replicación de Almacenamiento

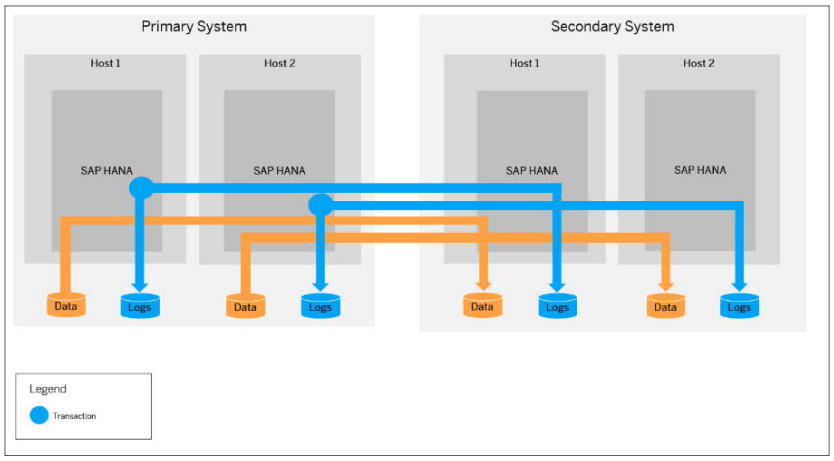


Fuente: SAP HANA Storage Replication, disponible en https://help.sap.com/viewer/6b94445c94ae495c83a19646e7c3fd56/2.0.03/en-US/2a3b86c65f0d485cb39ff10181986125.html, recuperado el 15 agosto 2018.

Replicación Del Sistema

La replicación de sistemas SAP HANA le permite configurar uno o más sistemas para que asuman el control de su sistema principal en escenarios de alta disponibilidad o recuperación ante desastres. Puede ajustar la replicación para satisfacer sus necesidades en términos de rendimiento y tiempo de conmutación por error.

Figura 47: Replicación del Sistema



Fuente: de SAP HANA System Replication, disponible en https://help.sap.com/viewer/6b94445c94ae495c83a19646e7c3fd56/2.0.03/en-US/b74e16a9e09541749a745f41246a065e.html, Recuperado el 15 de agosto de 2018.

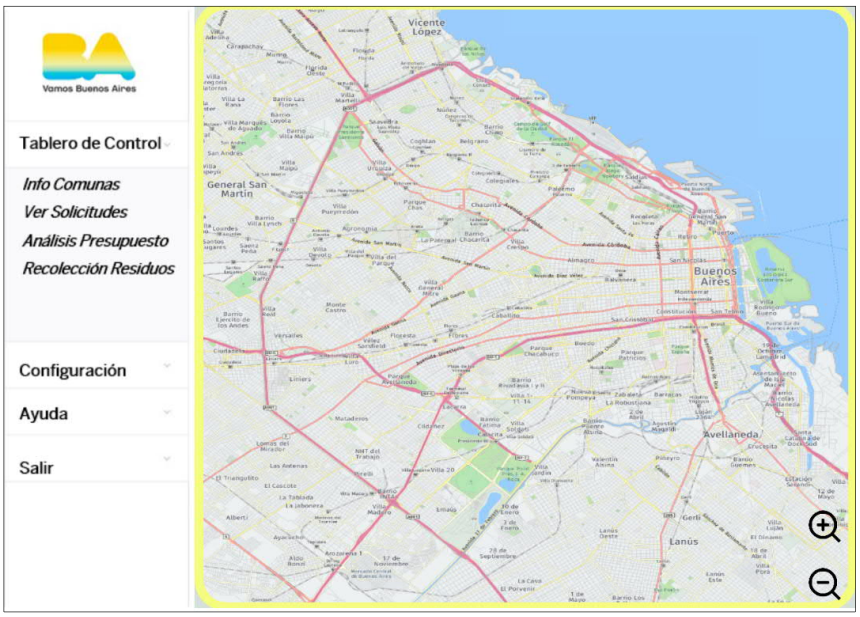
3.4 Aplicación Del Modelo / Diseño De Prototipo.

Con el objetivo de demostrar la funcionalidad integrada de los componentes de ML y GIS utilizados durante el análisis, se generó un prototipo que permite visualizar el funcionamiento del tablero de gestión. El prototipo está compuesto por:

3.4.1 Pantalla De Inicio

La pantalla de inicio se da comienzo con un mapa de la CABA conteniendo el menú para poder acceder a la funcionalidad del tablero de gestión.

Figura 48: Pantalla Principal - Mapa CABA



Fuente: Elaboración propia.

Las opciones disponibles son:

Info Comunas:

Muestra el detalle de la comuna con sus datos estadísticos, obtenidos desde los diversos orígenes de datos. Esta funcionalidad permite seleccionar una de las 15 comunas existentes en la CABA para realizar el análisis de forma individual.

Ver Solicitudes

Permite visualizar en el mapa las solicitudes existentes según el concepto seleccionado.

Análisis De Presupuesto

Muestra un gráfico con la información histórica del presupuesto en conjunto con los valores estimados según el resultado de los algoritmos de ML.

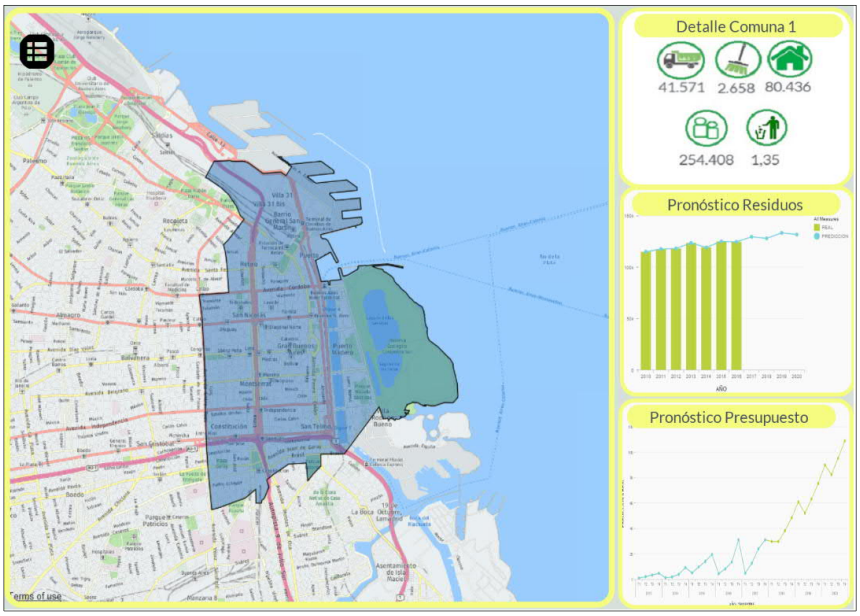
Recolección De Residuos

Muestra un gráfico que contienen los datos históricos de la recolección de residuos para la comuna seleccionada y el pronóstico futuro según el resultado de los algoritmos de ML.

3.4.2 Datos De La Comuna

Al seleccionar una comuna, se visualiza el polígono que la conforma junto con la información detallada de la misma, incluyendo los pronósticos de recolección de residuos y presupuesto ya calculados.

Figura 49: Representación gráfica de la Comuna 1



Fuente: Elaboración propia.

La pantalla se divide en 4 áreas:

Área Del Mapa De La Ciudad

En esta sección se muestra la información geográfica relacionada con la comuna seleccionada.

Detalle De Comuna

En esta sección se muestra el valor de datos de recolección de residuos para la comuna seleccionada. La descripción de cada uno de los iconos se detalla en la siguiente tabla:

Tabla 10: Significado de los Iconos en pantalla

	Cantidad en Tn de residuos destinados a relleno sanitario
	Cantidad de Residuos en Tn recolectados como resultado del barrido
	Cantidad de Residuos en Tn recolectados desde los hogares
	Cantidad de Habitantes para la comuna seleccionada
	Promedio de residuos en Tn por habitante

Fuente: Elaboración propia.

Pronóstico De Residuos

Este gráfico muestra la predicción de recolección de residuos para la comuna seleccionada en base a los datos y la configuración actual.

Pronóstico De Presupuesto

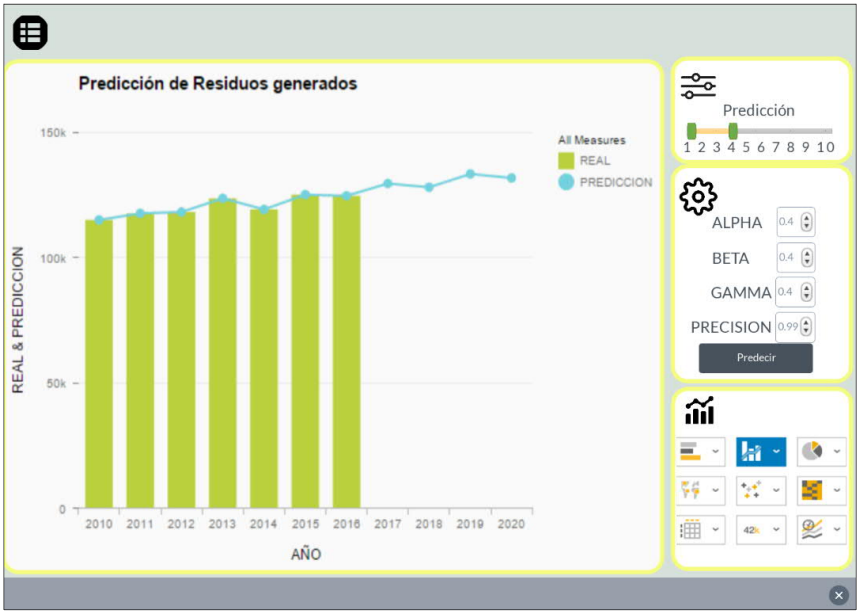
Este gráfico muestra la predicción del presupuesto requerido para la comuna seleccionada en base a los datos y a la configuración actual.

Seleccionando cualquiera de los gráficos de pronóstico el usuario puede visualizar el detalle del mismo.

3.4.3 Predicción De Residuos Generados (Detalle)

Este gráfico muestra la predicción de recolección de residuos para la comuna seleccionada en base a los datos y la configuración actual.

Figura 50: Predicción de Residuos Generados



Fuente: Elaboración propia.

Tabla 11: Opciones de Configuración ML

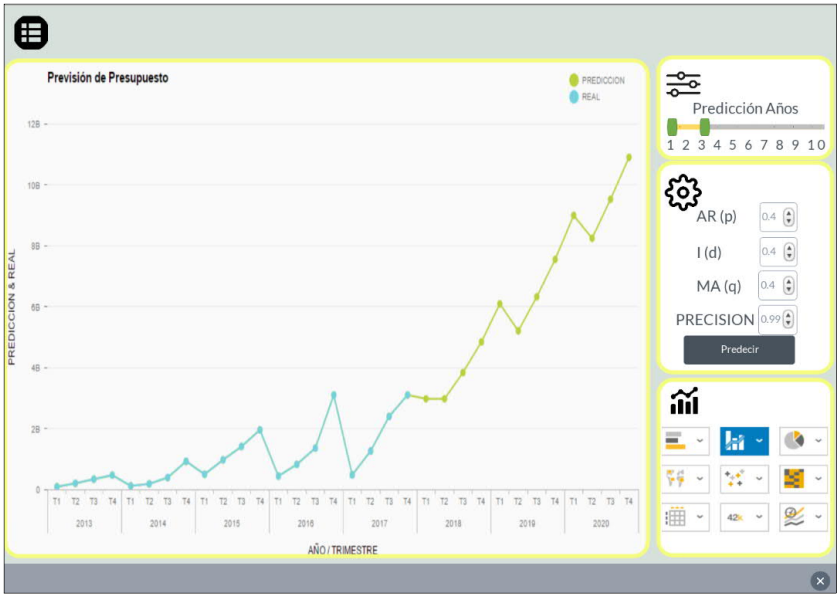
	Permite seleccionar la cantidad de meses a estimar
	Permite configurar los parámetros que se utilizarán en el algoritmo de ML
	Permite seleccionar el tipo de gráfico a visualizar

Fuente: Elaboración propia.

3.4.4 Previsión Del Presupuesto

Este gráfico muestra la predicción del presupuesto requerido para la comuna seleccionada en base a los datos y a la configuración actual.

Figura 51: Previsión del Presupuesto



Fuente: Elaboración propia.

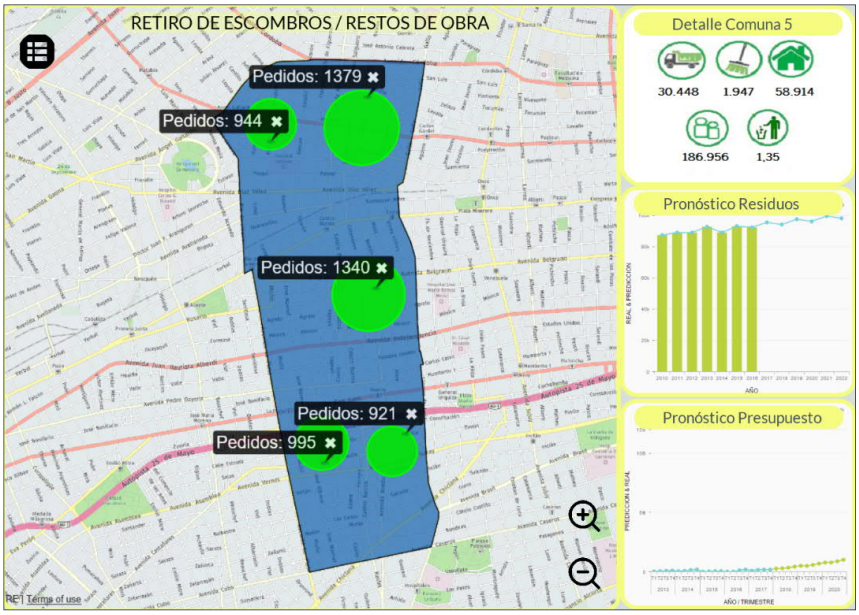
3.4.5 Visualizar Solicitudes

Esta opción permite visualizar en el mapa de la ciudad las solicitudes existentes agrupadas por concepto, según la configuración.

En la siguiente imagen se muestra el detalle de la Comuna 5 para las solicitudes bajo el concepto “RETIRO DE ESCOMBROS / RESTOS DE OBRA”.

Cada circulo representa un grupo generado por la función K-Means.

Figura 52: Visualización de Solicitudes



Fuente: Elaboración propia.

De esta forma el usuario puede analizar los diferentes conceptos relacionados, detectar focos, modificar los recorridos de los camiones de recolección, etc.

4. PROPUESTA DE IMPLEMENTACIÓN

4.1 Costo de Implementación

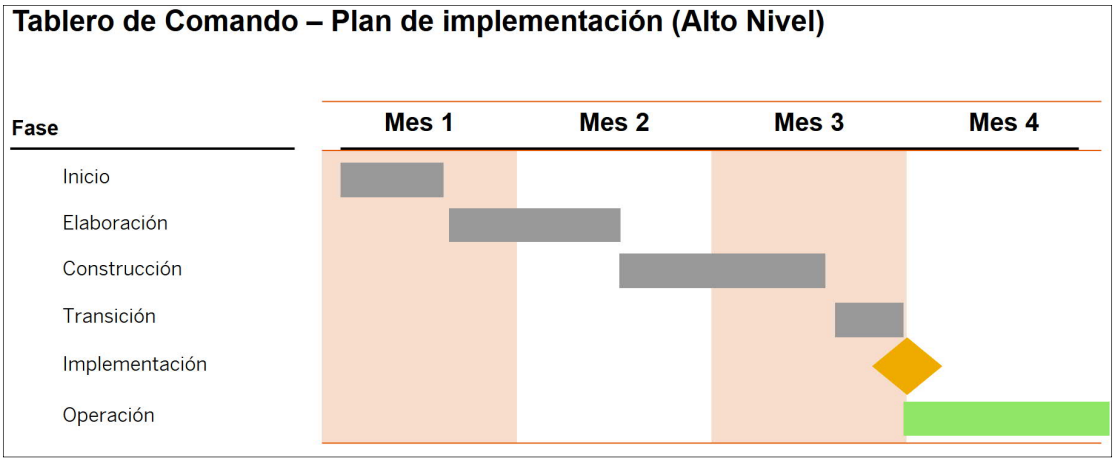
4.1.1 Propuesta de implementación

La implementación de la solución no requiere de una gran inversión a nivel de software, esto se debe a la utilización de software gratuito SAP HANA Express Edition, sobre Linux SLES 12.

Para el análisis de los costos de implementación del tablero de gestión, con la funcionalidad detallada se estima un tiempo de 3 meses para la implementación inicial, con un equipo reducido utilizando una metodología ágil.

El siguiente gráfico muestra un plan de implementación de 3 meses, con un mes de soporte a las operaciones y uso del tablero.

Figura 53 - Plan de implementación propuesto



Fuente: Elaboración Propia.

4.1.2 Equipo de Proyecto

Se estima un equipo reducido de trabajo que se encuentre dedicado a la implementación con una asignación mensual (expresada en días) que se detalla en la siguiente tabla:

Tabla 12 - Equipo de implementación

Tipo de Recurso	Mes 1	Mes 2	Mes 3	Mes 4	Total
Gerente de Proyecto	10	10	10	5	35
Responsable de Negocio	20	10	10	10	50
Analista de Datos	20	20	10	10	60
Programador	-	15	20	10	45
Administrador HANA	-	15	20	10	45
Total	50	70	70	45	235

Fuente: Elaboración propia.

El esfuerzo total de implementación estimado en días es de 235 días hombre.

Los roles necesarios para la implementación:

Gerente de Proyecto (PM): Responsable del éxito general del Proyecto. El PM debe autorizar y aprobar todos los gastos del proyecto. El PM también es responsable de aprobar que las actividades laborales cumplan con los criterios de aceptabilidad establecidos y se encuentren dentro de las variaciones aceptables. Se encargará de informar sobre el estado del proyecto de conformidad con el plan de gestión de las comunicaciones. Evaluará el desempeño de todos los miembros del

equipo del proyecto y comunicará su desempeño a los gerentes funcionales. El PM también es responsable de adquirir recursos humanos para el proyecto a través de la coordinación con los gerentes funcionales. El PM debe poseer las siguientes habilidades: liderazgo, gestión, planificación, control y comunicación efectiva.

Responsable de Negocio: Provee soluciones a los problemas que surgen en el curso de los negocios. Es el nexo principal entre el proyecto y el negocio. Debe revisar y validar los requerimientos, así como también validar que las métricas definidas se ajusten a la necesidad del negocio.

Es el responsable de dar la aprobación de las pruebas y autoriza la puesta en producción de los diferentes componentes del tablero de gestión.

Analista de Datos: Responsable de recopilar y analizar la información de las empresas para emitir recomendaciones o tomar decisiones. Debe definir en conjunto con el responsable de negocio los requerimientos y las métricas que se utilizarán en el tablero de gestión.

Es el responsable de definir los algoritmos estadísticos utilizados para el pronóstico, así como también validar los modelos de datos y asegurar el entrenamiento de los algoritmos de ML.

Programador: El programador es responsable de realizar las especificaciones técnicas, codificación y pruebas unitarias de los programas desarrollados.

Es el encargado de generar los paquetes de transporte que se importaran en los ambientes de calidad y producción.

Experto HANA: Es el encargado de administrar la plataforma SAP HANA, asegurando su operación, monitoreo y es el encargado de importar los paquetes de transporte de software.

Recursos de Hardware (CGP): La estimación se basa en el valor mensual informado por GCP para tres ambientes, desarrollo, calidad y producción necesarios para la implementación, según las mejores prácticas recomendadas por SAP.

En base al equipo de proyecto y el hardware necesario para la implementación se estima el siguiente costo total, detallado por mes:

Tabla 13: Costo de Implementación Estimado

Recursos	Costo Mensual ARS [13]	Mes 1	Mes 2	Mes 3	Mes 4	Total
Gerente de Proyecto	$18.240	$182.400	$182.400	$182.400	$91.200	$638.400
Responsable de Negocio	$15.200	$304.000	$152.000	$152.000	$152.000	$760.000
Analista de Datos	$15.200	$304.000	$304.000	$152.000	$152.000	$912.000
Programador	$12.160	$0	$182.400	$243.200	$121.600	$547.200
Administrador HANA	$12.160	$0	$182.400	$243.200	$121.600	$547.200
Total		$790.400	$1.003.200	$972.800	$638.400	$3.404.800

Fuente: Elaboración Propia.

4.1.3 Costo del Hardware

Durante la implementación del tablero de gestión, se puede ir aprovisionando los servidores necesarios de forma escalonada, según el plan propuesto.

La siguiente tabla muestra un detalle de dicha necesidad por cada mes:

[13] El costo mensual se calcula en base al estimado proporcionado por el proveedor GCP de $213 dólares mensuales a un tipo de cambio de 1 Dólar = 38 Pesos Argentinos.

Tabla 14: Detalle del Hardware requerido por Mes

Hardware (GCP)	Mes 1	Mes 2	Mes 3	Mes 4
Servidor Desarrollo	-	1	1	1
Servidor Calidad	-	-	1	1
Servidor Producción	-	-	1	1
Total		1	3	3

Fuente: Elaboración Propia.

De esta forma y gracias a la velocidad de rápido aprovisionamiento de GCP el ambiente de desarrollo puede habilitarse recién en el segundo mes de proyecto, mientras que los servidores de Calidad y producción lo harán únicamente en el tercer mes.

El costo desglosado por mes, para provisionar los servidores en GCP, según la estimación proporcionada por el proveedor es la siguiente:

Tabla 15: Costo de Hardware

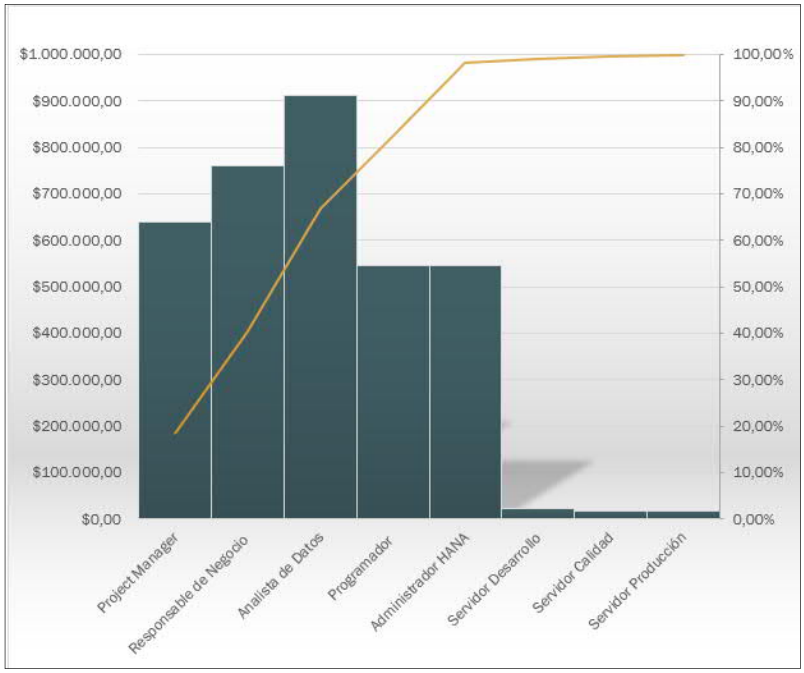
Hardware (GCP)	Costo Diario (Pesos)	Mes 1	Mes 2	Mes 3	Mes 4	Total
Servidor Desarrollo	$8.102	$0	$8.102	$8.102	$8.102	$24.305
Servidor Calidad	$8.102	$0	$0	$8.102	$8.102	$16.203
Servidor Producción	$8.102	$0	$0	$8.102	$8.102	$16.203
Total		$0	$8.102	$24.305	$24.305	$56.711

Fuente: Elaboración Propia.

4.1.4 Análisis de Costos de implementación

El siguiente gráfico muestra un análisis de Pareto de distribución de costos estimado para la implementación.

Figura 54: Análisis de Costo



Fuente: Elaboración Propia.

4.1.5 Distribución de costos

La siguiente tabla muestra la distribución de los costos de implementación:

Tabla 16: Distribución de costos

Descripción	Costo Total	Porcentaje del Total	Porcentaje Acumulado
Gerente de Proyecto	$638.400,00	18,44%	18,44%
Responsable de Negocio	$760.000,00	21,96%	40,40%
Analista de Datos	$912.000,00	26,35%	66,75%
Programador	$547.200,00	15,81%	82,55%
Administrador HANA	$547.200,00	15,81%	98,36%
Servidor Desarrollo	$24.304	5,04%	99,06%
Servidor Calidad	$16.203	3,36%	99,53%
Servidor Producción	$16.203	3,36%	100,00%
Total	$3.461.511,20	100,00%	

Fuente: Elaboración Propia.

4.1.6 Mantenimiento del Hardware

La estimación de mantenimiento de los servidores en GCP, desglosado por día/mes/año, es el siguiente:

Tabla 17: Costo de Mantenimiento de Hardware

Hardware (GCP)	Costo Mensual (USD)	T/C	Diario	Mensual	Anual
Servidor Desarrollo	$213	$38	$270	$8.102	$97.219
Servidor Calidad	$213	$38	$270	$8.102	$97.219
Servidor Producción	$213	$38	$270	$8.102	$97.219
Total			$810	$24.305	$291.658

Fuente: Elaboración Propia.

4.1.7 Ahorro Estimado

Si bien la implementación del tablero de gestión no implica un ahorro para el Gobierno de la Ciudad, se espera que gracias a la implementación de este se pueda mejorar la toma de decisiones, así como también a la gestión de los indicadores definidos.

El principal beneficio esperado es la mejora de la opinión pública, por ejemplo, en el manejo de los fondos públicos. Para ello será necesario definir los KPIs relevantes y definir un proceso de medición orientado a estos KPIs. Dichos KPIs y procesos no están definidos en el momento en que se realizó este trabajo, por lo que no pudieron ser evaluados durante la elaboración de esta tesis.

4.1.8 Beneficios Esperados

Se espera que la utilización del tablero de gestión propuesto provea los siguientes beneficios al Gobierno de la Ciudad

- Implementar una plataforma en memoria que brinde flexibilidad y herramientas para facilitar el proceso de transformación digital.
- Generar información predictiva en base a datos históricos con un alto poder de precisión.
- Dar una visibilidad temprana sobre la evolución de la generación de residuos en la CABA, que permita tomar acciones proactivas para alcanzar el objetivo de Basura Cero planteado en la ciudad.
- Permitir estimar el presupuesto necesario para los próximos períodos que permitan dar transparencia y mejorar la eficiencia en la planificación.
- Lograr una eficiencia en la planificación de presupuesto permitirá mejorar la opinión pública sobre la gestión de fondos del estado.
- Simplificar la gestión de solicitudes al SUACI mediante el procesamiento de las solicitudes geográficamente relacionadas.
- Comparar los indicadores de las diferentes comunas con el fin de identificar aquellas que logran mejores resultados, con el fin de analizar aquellas acciones que puedan ser utilizadas en otras comunas.

4.1.9 Mejoras a futuro

La transformación digital está alcanzando a todas las compañías, una elección correcta de las herramientas para lograrla puede convertirse en una ventaja competitiva, para formar los fundamentos de una Ciudad Inteligente.

Este tipo de iniciativa requiere que IT pueda brindar tiempos de reacción acordes a las necesidades de la compañía, en general no superior a 3 meses, para lograr implementar una solución innovadora con capacidad de ser adaptada y mejorada continuamente.

La utilización del tablero de gestión en CABA permitirá al Gobierno de La Ciudad de Buenos Aires contar con una nueva herramienta para analizar algunos de los temas que están dentro de sus intereses y de la población: El objetivo de Basura Cero, La recolección de residuos en la Ciudad, en especial los residuos Voluminosos, y la planificación eficaz del presupuesto destinado a la recolección de residuos.

Podríamos decir que esta es la punta del iceberg, y que es el comienzo de un camino de transformación digital en el que queda mucho por recorrer.

La implementación de una plataforma en memoria permite no solo utilizar las capacidades de Analíticas y de Machine Learning, sino que es la puerta a la incorporación de otras tecnologías disponibles y que gracias a la utilización de SAP HANA están disponibles.

Existen diversos usos a futuro que pueden planificarse para el tablero de gestion propuesto, tanto a nivel de funcionalidad, como de alcance geográfico.

Funcionalidad:

Origen de los Datos: El prototipo propuesto se basa en datos estáticos obtenidos de diversas fuentes de datos públicas. La unificación de la fuente de datos en una única plataforma permitirá procesar los datos para obtener información que dé un valor agregado al negocio, permitiendo también incorporar nuevos orígenes de datos. La gestión integral de los datos es uno de los fundamentos para poder optimizar el proceso de la información y maximizar su potencial.

Internet de las Cosas (IoT): El análisis de la información de solicitudes al SUACI es solo una parte de la historia. El proceso de recolección de residuos contiene muchas variables de las que hoy no contamos en su totalidad, como se la ubicación de los camiones, rutas de recolección, estado de los camiones recolectores, entre otras.

El uso de IoT mediante la utilización de GPS, y sensores permitiría recolectar información y estado de los mismos y de esta forma nutrir de información a los sistemas de información.

Rutas de recolección: La ubicación geográfica de las solicitudes al SUACI, permite generar una ruta de recolección óptima para el día, optimizando los tiempos y costos de recolección de residuos voluminosos.

Análisis en Tiempo real: Una de las ventajas de las plataformas en memoria es la capacidad de procesamiento en tiempo real. Una vez unificada la fuente de datos, utilizando las capacidades de IoT, y la incorporación de rutas de recolección, será posible la el procesamiento de los datos en tiempo real. Con esta funcionalidad se podrá lograr, por ejemplo, un seguimiento de las solicitudes al SUACI en el momento que son creadas, permitiendo modificar en tiempo real la ruta de los camiones de recolección cercanos, si los mismos están en condiciones de cumplir esa solicitud. No dejes para mañana lo que puedes hacer hoy, dejará de ser un dicho para ser una realidad.

Nuevas funciones analíticas: El tablero de gestión propuesto analiza las series de tiempo existentes y genera información predictiva con un alto nivel de precisión en base a las mismas, la plataforma SAP HANA posee un amplio repositorio de funciones

predictivas que pueden ser utilizadas para sumar valor al tablero de gestion, principalmente las funciones de Clasificación (ej. Arboles de decisión), Regresión (Ej. Regresión Linear) y Grafos (cálculo de rutas de recolección).

Simulación / Retroalimentación Los escenarios de Análisis predictivo y Machine Learning tienen la capacidad de retroalimentarse con las variables y actualizar sus modelos. Incluir un proceso de retroalimentación, o actualización en tiempo real de los datos, permite a los algoritmos refinar sus resultados y/o simulación de los resultados en base a cambios en las variables.

Alcance Geográfico:

La solución presentada se centró en la Ciudad de Buenos Aires, aunque no hay limitación geográfica para el uso de los indicadores propuestos, siendo posible utilizar a nivel Provincial o Nacional, con tiempos de implementación y alcances similares.

5. CONCLUSIONES

La estrategia propuesta para la recolección de residuos por el Gobierno de la Ciudad de Buenos Aires es ambiciosa, y requiere de una planificación y seguimiento continua. El uso de herramientas Analíticas y de Machine Learning permiten integrar y explotar la información existente de una forma que antes no era posible. Durante este estudio se detectaron algunos puntos que pueden ser mejorados:

Como resultado de la predicción de recolección de residuos podemos anticipar que la meta de Basura Cero, propuesta para el año 2020, muy difícilmente sea cumplida, a menos que se tomen medidas correctivas para modificar la tendencia actual en la generación de los residuos.

El análisis del consumo del presupuesto planificado, en relación con el consumido, muestra desviaciones que se incrementan cada año. La utilización de funciones predictivas brinda una herramienta adicional para soportar la toma de decisiones. De esta forma se espera poder anticipar los valores futuros y tener una planificación más exacta.

El uso de la funcionalidad geográfica y la posibilidad de visualizar en el mapa de la ciudad las diferentes solicitudes, permite rápidamente detectar focos de atención gracias a la agrupación realizada por el algoritmo. Esto provee una nueva fuente de información en base a datos ya existentes y permite explorar la misma de forma dinámica.

El proceso de los datos, no relacionados y proveniente de diferentes orígenes públicos, luego integrados en una única plataforma, permiten generar información útil que no está disponible a simple vista, generando un valor agregado gracias a la tecnología y herramientas de Machine Learning.

Finalmente, el uso de SAP HANA permite procesar los datos rápidamente, gracias a sus capacidades de BD en memoria, y utilizar una gran variedad de algoritmos y funcionalidades nativos de la plataforma para procesar la información.

Principalmente la utilización de las tecnologías en conjunto permite agregar valor para la toma de decisiones. Se espera que la utilización de estas tecnologías sean el punto inicial para incorporar estas herramientas a otros ámbitos, ya sea a nivel nacional o dentro de otras áreas de la ciudad.

6. Referencias

Abadi, D. J., Madden, S. R., & Ferreira, M. C. (2006). Integrating Compression and Execution in Column-Oriented Database Systems. *SIGMOD 2006, June 27–29, 2006.* Chicago, Illinois, USA.

Abadi, D. J., Madden, S. R., & Hachem, N. (2008). Column-Stores vs. Row-Stores: How Different Are They. *SIGMOD'08, June 9–12, 2008.* Vancouver, BC, Canada.

Artho, P. R. (30 de June de 2014). *In-Memory Databases State of the Art of In-Memory Databases.* Seminar "Advanced Database Systems"- Master of Science in Engineering - University of Applied Sciences Rapperswil.

Bernabeu, R. D. (2010). *DATA WAREHOUSING: Investigación y Sistematización de Conceptos.* Córdoba, Córdoba , Argentina: Hefesto. Recuperado el 12 de 05 de 2018

Boncz, P., Manegold, S., & Kersten, M. (1999). Database Architecture Optimized for the new Bottleneck: Memory Access. *25th VLDB Conference.* Edinburgh, Scotrland.

Chee, T., Chan, L.-K., Chuah, M.-H., Tan, C.-S., Wong, S.-F., & Yeoh, W. (2009). *Business Intelligence Systems: State-of-the-art review.* (S. o. Technology, Ed.) Recuperado el 13 de 05 de 2018

Cysco Systems. (2009). *Cisco cloud computing Data Center Strategy, Architecture and Solutions.* . Recuperado el 12 de 05 de 2018, de https://www.cisco.com/c/dam/en_us/solutions/industries/docs/gov/CiscoCloudComputing_WP.pdf

Dario, B. R. (2010). *DATA WAREHOUSING: Investigación y Sistematización de Conceptos.* Córdoba , , Córdoba , Argentina.

Davenport , H. T., & Harris, G. J. (2007). *Competing on analytics: The new science of Winning.* Harvard Business Press. Obtenido de https://books.google.com.ar/books?hl=es&lr=&id=n7Gp7Q84hcsC&oi=fnd&pg=PR4&ots=9uVJIQ-AOH&sig=wN9M1IORiaxWOh4xoGF2XKHrloQ#v=onepage&q&f=false

Davenport, H. T. (2009). Make Better Decisions. *Hardware Business Review*, 9.

Estadísticas y Censos. (2016). *Residuos recolectados por tipo y promedio diario por habitante. Ciudad de Buenos Aires. Años 1995/2016*. (Buenos Aires Ciudad) Recuperado el 08 de 2018, de http://www.estadisticaciudad.gob.ar/eyc/?p=29140

Garcia Molina, H. (1992). Main Memory Database Systems: An Overview. *IEEE Transactions on knowledge and Data engineering, Vol 4, NO 6* . Stanford, CA.

Gartner. (s.f.). *Gartner IT Glossary*. Recuperado el 28 de 97 de 2018, de https://www.gartner.com/it-glossary/business-analytics

Gibson, M., Arnott, D., & Jagielska, I. (2004). Evaluating the Intangible Benefits of Business Intelligence: Review & Research Agenda. *IFIP International Conference on Decision Support Systems (DSS2004)*.

Gobierno de la Ciudad de Buenos Aires. (2007). *DECRETO N° 639/07*. CIUDAD AUTÓNOMA DE BUENOS AIRES: TELERMAN - VELASCO - BEROS. Recuperado el 22 de 7 de 2018, de http://www.buenosaires.gob.ar/areas/leg_tecnica/sin/normapop09.php?id=98735&qu=h&ft=0&cp=&rl=0&rf=0&im=&ui=0&printi=&pelikan=1&sezion=&primera=0&mot_toda=&mot_frase=&mot_alguna=&digId=

Gobierno de la Ciudad de Buenos Aires. (Diciembre de 2017). *Buenos Aires Data – Gobierno de la Ciudad*. Obtenido de https://data.buenosaires.gob.ar/

Gobierno de La Ciudad de Buenos Aires. (2017). *Buenos Aires-Ambiente y Espacio Público-Higiene-Recolección*. Recuperado el 26 de 08 de 2018, de http://www.buenosaires.gob.ar/ambienteyespaciopublico/higiene/recoleccion

Gobierno de la Ciudad de Buenos Aires. (s.f.). *Estadisticas y Censos*. (Dirección General de Estadística y Censos) Recuperado el Diciembre de 2017, de https://www.estadisticaciudad.gob.ar/eyc/

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton , New Jersey: Princeto University Press.

Harrington, P. (2012). *Machine Learning in Action.* Manning.

Hoornweg, D., & Bhada-Tata, P. (2012). *WHAT A WASTE - A Global Review of Solid Waste Management.* Washington,: World Bank.

Hossain, S., Islam, F., Karim, R., & Siddique, K. N.-E.-A. (2014). *A Critical Comparison Between Distributed Database Approach and Data Warehousing.*

Inmon, W. (2002). *Building the Data Warehouse*. Robert Ipsen. doi:ISBN: 0-471-08130-2

Kalekar, P. S. (2004). Time series Forecasting using Holt-Winters Exponential Smoothing. Kanwal Rekhi School of Information Technology.

Kaplan, R. (1999). *The Balanced Scorecard for Public-Sector Organizations.* Harvard Business Review.

Kaplan, R., & Norton, D. (1997). *The Balanced Scorecard: Translating Strategy into Action* (vol. 85, no 9, p. 1509-1510. ed.). Proceedings of the IEEE.

Kimball, R., & Ross, M. (2013). *The Data Warehoyse Toolkit* (3ra ed.). Wiley.

Legislatura de la Ciudad Autónoma de Buenos Aires. (24 de 11 de 2005). LEY Nº 1854/05. *LEY Nº 1854/05*. Buenos Aires, CABA, Argentina: LEGISLATURA DE LA CIUDAD AUTÓNOMA DE BUENOS AIRES. Obtenido de http://www.buenosaires.gob.ar/areas/leg_tecnica/sin/normapop09.php?id=81508&qu=c&cp&rl=1&rf&im&mot_toda&mot_frase&mot_alguna

Luhn, H. P. (Oct de 1958). *A Business Intelligence System.* (v. 2. IBM Journal of Research and Development, Ed.) doi:10.1147/rd.24.0314

Mell, P., & Grance, T. (2011). The NIST definition of Cloud Computing. *NIST Special Publication 800-145*. US: National Institute of Standards and Technology.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Odriozola, V. (2014). *Plan de Basura Cero para Buenos aires.* Buenos aires: Greenpeace.

Plattner, H. (2009). A common Database Approach for OLTP and OLAP Using an In-Memory Column Database. *SIGMOD 09.* Providence, Rhode Island, USA: Hasso Plattner Institute for IT Systems Engineering.

Plattner, H. (2014). The Impact of Columnar In-Memory Databases on Enterprise Systems. *Hasso Plattner Institute for IT Systems Engineering*. Potsdam, Germany: Hasso Plattner Institute for IT Systems Engineering.

Power, D. J. (2007). *A brief history of decision support systems*, 4.1. Recuperado el 10 de 8 de 2018, de http://dssresources.com/history/dsshistory.html

Qi Zhang, Lu Cheng, & Raouf Boutaba. (2010). Cloud computing: state-of-the-art and research challenges. The Brazilian Computer Society 2010.

Rimal, B. P., Jukan, A., Katsaros, D., & Goeleven, Y. (2010). Architectural Requirements for Cloud Computing Systems: An Enterprise Cloud Approach. Springer Science+Business Media B.V. 2010.

SAP SE. (2017). Frequently Asked Questions AboutSAP HANA, Express Edition. N/A: SAP SE. Recuperado el 17 de 7 de 2018, de https://www.sap.com/documents/2016/09/88baee6d-897c-0010-82c7-eda71af511fa.html

SAP SE. (s.f.). *SAP HANA Predictive Analysis Library (PAL)*. Recuperado el 12 de 08 de 2018, de SAP HELP Portal: https://help.sap.com/viewer/2cfbc5cf2bc14f028cfbe2a2bba60a50/2.0.02/en-US/c9eeed704f3f4ec39441434db8a874ad.html

SAP SE. (s.f.). *SAP HANA Spatial Reference*. Recuperado el 02 de 09 de 2018, de SAP HELP Portal: https://help.sap.com/viewer/cbbbfc20871e4559abfd45a78ad58c02/2.0.02/en-US/e1c934157bd14021a3b43b5822b2cbe9.html

SAP SE. (s.f.). *SAP Productos - HANA*. Recuperado el 02 de 09 de 2018, de SAP: https://www.sap.com/latinamerica/products/hana.html

Seeger, M. (2002). Matthias Seeger. Learning with labeled and unlabeled data. Institute for Adaptive and Neural Computation University of Edinburgh.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2002). *Fundamento de Bases de Datos* (Cuarta Edición ed.). Madrid: McGRAW-HILL/INTERAMERICANA DE ESPAÑA, S. A. U.

Upadhyay, R. (s.f.). *ARIMA Models – Manufacturing Case Study Example (Part 3)*. Recuperado el 12 de 08 de 2018, de http://ucanalytics.com/blogs/arima-models-manufacturing-case-study-example-part-3/

Wikipedia. (s.f.). *Juegos de guerra*. Obtenido de https://es.wikipedia.org/wiki/Juegos_de_guerra#Trama

Wikipedia. (s.f.). *Modelo autorregresivo integrado de media móvil*. Recuperado el 12 de 08 de 2018, de https://es.wikipedia.org/wiki/Modelo_autorregresivo_integrado_de_media_m%C3%B3vil