Universidad Técnica Nacional

Minería de Datos

ISW 911

Proyecto #1

Alexander Cruz Arroyo

Jorge Porras Araya

03/03/2020

Contenido

[Introducción 3](#_Toc34162244)

[Justificación 3](#_Toc34162245)

[Objetivo General 4](#_Toc34162246)

[Objetivos Específicos 4](#_Toc34162247)

[Antecedentes 5](#_Toc34162248)

[Base de datos 5](#_Toc34162249)

[Data Warehouse 5](#_Toc34162250)

[Big Data 6](#_Toc34162251)

[Minería de Datos 6](#_Toc34162252)

[Procesos ETL 7](#_Toc34162253)

[Fase de Extracción en los procesos ETL 8](#_Toc34162254)

[Procesos ETL: fase de Transformación 8](#_Toc34162255)

[Proceso de Carga: la culminación de los procesos ETL 8](#_Toc34162256)

[Alcance 9](#_Toc34162257)

[Power BI 9](#_Toc34162258)

[SQL Server Management Studio 9](#_Toc34162259)

[Visual Studio 9](#_Toc34162260)

[Conclusiones 10](#_Toc34162261)

[Referencias 11](#_Toc34162262)

# Introducción

En este proyecto se pusieron en práctica los conocimientos adquiridos de las técnicas y herramientas analizadas en el curso de minería de datos, de tal manera que pueda aportar información útil a la gerencia para la toma de decisiones.

Se analizaron los datos de una compañía eléctrica que trabaja con medidores de energía ION, los cuales pueden llegar a mal funcionar y presentar ciertos tipos de eventos y fallos distintos, ubicados en diferentes sedes de Costa Rica. Se cuenta con una base de datos con unos 650 mil registros aproximadamente, sobre las sedes, las diferentes causas de fallo, tipos de eventos, fechas, etc… Para la cual se creó un datawarehouse, modelos y procesos para poder brindar información importante a la gerencia, sobre los datos existentes en la base de datos.

# Justificación

La expresión cualitativa sobre el por qué se decidió realizar este proyecto sobre dicha base de datos, sale de la necesidad de ordenar, priorizar y analizar toda la información contenida en la BD acerca de los medidores ubicados en las distintas sedes de la compañía alrededor del país, esto con la posibilidad de seguir analizando la información en un corto, mediano y largo plazo, actualizándose continuamente uniformemente con la información, gracias a los modelos y plantillas de estadísticas realizadas.

# Objetivo General

Analizar los datos contenidos en la base de datos de la compañía, para visualizar porcentajes y resultados a partir de esta información, determinar ciertos comportamientos, fallos anormales o muy repetitivos dependiendo de las zonas, como también muchas más estadísticas con el fin de mejorar y hacer crecer la compañía mediante al análisis de todos los datos.

# Objetivos Específicos

* Crear estrategias para brindar la información a la gerencia.
* Generar estadísticas sobre los registros presentes en el BD.
* Determinar comportamientos irregulares por medio del análisis de datos.
* Realizar mejoras en la empresa en base a la información analizada.
* Practicar los conocimientos adquiridos sobre las técnicas y herramientas analizadas.

# Antecedentes

Vamos a iniciar con los conceptos básicos para quien quiera ver y entender este proyecto pueda hacerlo, definiremos que son las bases de datos, data warehouse, minería de datos y algunos conceptos más, para una comprensión amplia del tema.

## Base de datos

Una base de datos es un “almacén” que nos permite guardar grandes cantidades de información de forma organizada para que luego podamos encontrar y utilizar fácilmente.

El término de bases de datos fue escuchado por primera vez en 1963, en un simposio celebrado en California, USA. Una base de datos se puede definir como un conjunto de información relacionada que se encuentra agrupada ó estructurada.

Desde el punto de vista informático, la base de datos es un sistema formado por un conjunto de datos almacenados en discos que permiten el acceso directo a ellos y un conjunto de programas que manipulen ese conjunto de datos.

Cada base de datos se compone de una o más tablas que guarda un conjunto de datos. Cada tabla tiene una o más columnas y filas. Las columnas guardan una parte de la información sobre cada elemento que queramos guardar en la tabla, cada fila de la tabla conforma un registro.

## Data Warehouse

El concepto de data warehouse se originó en 1988 con el trabajo de los investigadores de IBM, Barry Devlin y Paul Murphy aunque el término data warehouse fue acuñado por William H. Inmon, el cual es conocido como el padre de Data Warehousing. Inmon describió un data warehouse como una colección de datos orientada a un tema específico, integrado, variante en el tiempo y no volátil, que soporta el proceso de toma de decisiones.

Un data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso.

Normalmente, una data warehouse se aloja en un servidor corporativo o cada vez más, en la nube. Los datos de diferentes aplicaciones de procesamiento de transacciones Online (OLTP) y otras fuentes se extraen selectivamente para su uso por aplicaciones analíticas y de consultas por usuarios.

Data Warehouse es una arquitectura de almacenamiento de datos que permite a los ejecutivos de negocios organizar, comprender y utilizar sus datos para tomar decisiones estratégicas. Un data warehouse es una arquitectura conocida ya en muchas empresas modernas.

## Big Data

Para entender qué significa realmente "big data", resulta útil conocer ciertos antecedentes históricos. A continuación, ofrecemos la definición de Gartner, de aproximadamente 2001 (y que continúa siendo la definición de referencia): Big data son datos que contienen una mayor variedad y que se presentan en volúmenes crecientes y a una velocidad superior. Esto se conoce como "las tres V".

Dicho de otro modo, el big data está formado por conjuntos de datos de mayor tamaño y más complejos, especialmente procedentes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede gestionarlos. Sin embargo, estos volúmenes masivos de datos pueden utilizarse para abordar problemas empresariales que antes no hubiera sido posible solucionar.

## Minería de Datos

La minería de datos es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Empleando una amplia variedad de técnicas, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más.

El proceso de hurgar en los datos para descubrir conexiones ocultas y predecir tendencias futuras tiene una larga historia. Conocido algunas veces como "descubrimiento de conocimientos en bases de datos", el término "minería de datos" no se acuño sino hasta la década de 1990. Pero su base comprende tres disciplinas científicas entrelazadas: estadística (el estudio numérico de relaciones de datos), inteligencia artificial (inteligencia similar a la humana exhibida por software y/o máquinas) y machine learning (algoritmos que pueden aprender de datos para hacer predicciones). Lo que era antiguo es nuevo otra vez, ya que la minería de datos continúa evolucionando para igualar el ritmo del potencial sin límites del big data y poder de cómputo asequible.

¿Entonces por qué es importante la minería de datos? Ha podido apreciar los números asombrosos – el volumen de datos producidos se duplica cada dos años. Los datos no estructurados por sí solos conforman el 90% del universo digital. Pero más información no significa necesariamente más conocimientos.

La minería de datos le permite:

* Filtrar todo el ruido caótico y repetitivo en sus datos.
* Entender qué es relevante y luego hacer un buen uso de esa información para evaluar resultados probables.
* Acelerar el ritmo de la toma de decisiones informadas.

# Procesos ETL

Los procesos ETL son una parte de la integración de datos, pero es un elemento importante cuya función completa el resultado de todo el desarrollo de la cohesión de aplicaciones y sistemas.

La palabra ETL corresponde a las siglas en inglés de:

• Extraer: extract.

• Transformar: transform.

• Y Cargar: load.

Con ello, queremos decir que todo proceso ETL consta precisamente de estas tres fases: extracción, transformación y carga. Vamos a definir en qué consisten cada una de estas fases.

### Fase de Extracción en los procesos ETL

Para llevar a cabo de manera correcta el **proceso de extracción, primera fase de los procesos ETL, hay** que seguir los siguientes pasos:

* Extraer los datos desde los sistemas de origen.
* Analizar los datos extraídos obteniendo un chequeo.
* Interpretar este chequeo para verificar que los datos extraídos cumplen la pauta o  
  estructura que se esperaba. Si no fuese así, los datos deberían ser rechazados.
* Convertir los datos a un formato preparado para iniciar el proceso de transformación.

Además, uno de las prevenciones más importantes que se deben tener en cuenta durante el proceso de extracción sería el **exigir siempre que esta tarea cause un impacto mínimo en el sistema de origen**. Este requisito se basa en la práctica ya que, si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que no pudiera volver a ser utilizado con normalidad para su uso cotidiano.

### Procesos ETL: fase de Transformación

**La fase de transformación de los procesos de ETL aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados**. Estas directrices pueden ser declarativas, pueden basarse en excepciones o restricciones, pero, para potenciar su pragmatismo y eficacia, hay que asegurarse de que sean:

* Declarativas.
* Independientes.
* Claras.
* Inteligibles.
* Con una finalidad útil para el negocio.

### Proceso de Carga: la culminación de los procesos ETL

En esta fase, los datos procedentes de la fase anterior (fase de transformación) son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes.

# Alcance

La meta final del proyecto es poder cumplir con todos los objetivos que la compañía o nosotros mismos impongamos, a la hora de visualizarlo terminado o bien ya en su máximo desarrollo, simplemente actualizándose conforme cambie la información, es por eso que, para poder realizar exitosamente todas las metas y objetivos, se trabajó con una serie de conceptos que ya conocimos y algunas herramientas que a continuación se darán a conocer un poco más de ellas.

## Power BI

Power BI es una solución de análisis empresarial que le permite visualizar sus datos y compartir conocimientos en toda su organización, o incrustarlos en su aplicación o sitio web. Se puede conectar con cientos de fuentes de datos y dar vida a los datos con paneles e informes en vivo.

## SQL Server Management Studio

SQL Server Management Studio (SSMS) es un entorno integrado para administrar cualquier infraestructura SQL. Use SSMS para acceder, configurar, administrar, administrar y desarrollar todos los componentes de SQL Server, Azure SQL Database y SQL Data Warehouse. SSMS proporciona una única utilidad integral que combina un amplio grupo de herramientas gráficas con varios editores de scripts enriquecidos para proporcionar acceso a SQL Server para desarrolladores y administradores de bases de datos de todos los niveles de habilidad.

## Visual Studio

Visual Studio es un conjunto de herramientas y otras tecnologías de desarrollo de software basado en componentes para crear aplicaciones eficaces y de alto rendimiento, permitiendo a los desarrolladores crear sitios y aplicaciones web, así como otros servicios web en cualquier entorno que soporte la plataforma.

En palabras más específicas, Visual Studio es un conjunto completo de herramientas de desarrollo para la generación de aplicaciones web ASP.NET, Servicios Web XML, aplicaciones de escritorio y aplicaciones móviles. Visual Basic, Visual C# y Visual C++ utilizan todos el mismo entorno de desarrollo integrado (IDE), que habilita el uso compartido de herramientas y facilita la creación de soluciones en varios lenguajes. Asimismo, dichos lenguajes utilizan las funciones de .NET Framework, las cuales ofrecen acceso a tecnologías clave para simplificar el desarrollo de aplicaciones web ASP y Servicios Web XML.

# Conclusiones

* Se brinda la información necesaria por medio de las estrategias adecuadas.
* El análisis de datos es efectivo a la hora ilustrar la información.
* Se determinaron ciertos patrones gracias al análisis de datos, que ayudan a mejorar la empresa.
* Se puede tener un mejor control de eventos o fallos gracias a la Informacion analizada y así evitar que sucedan.
* La determinación de ciertos comportamientos por sede o lugar, nos ayuda ser más eficientes.
* Se pudo crear un repositorio unificado para todos los datos.
* El análisis en Power BI, permite visualizar los datos y compartir conocimientos en toda la organización de una manera sencilla y eficaz.

# Referencias

1. Damián Pérez. (2007). ¿Qué son las bases de datos?. 03/03/2020, de maestros del web Sitio web: <http://www.maestrosdelweb.com/que-son-las-bases-de-datos/>
2. powerdata. (2019). Data Warehouse: todo lo que necesitas saber sobre almacenamiento de datos. 03/03/2020, de powerdata Sitio web: <https://www.powerdata.es/data-warehouse>
3. oracle. (2019). Definición de big data. 03/03/2020, de oracle Sitio web: <https://www.oracle.com/es/big-data/guide/what-is-big-data.html>
4. sas. (2019). Minería de datos. 03/03/2020, de sas Sitio web: <https://www.sas.com/es_mx/insights/analytics/data-mining.html>
5. powerdata. (2017). ¿Qué son los procesos ETL?. 03/03/2020, de powerdata Sitio web: <https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/qu-son-los-procesos-etl>
6. microsoft. (2020). What is Power BI?. 03/03/2020, de microsoft Sitio web: <https://powerbi.microsoft.com/en-us/what-is-power-bi/>
7. microsoft. (2019). What is SQL Server Management Studio. 03/03/2020, de microsoft Sitio web: <https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver15>
8. msn. (2020). Qué es y para qué sirve Visual Studio. 03/03/2020, de msn Sitio web: <https://www.msn.com/es-cl/noticias/microsoftstore/%C2%BFqu%C3%A9-es-y-para-qu%C3%A9-sirve-visual-studio-2017/ar-AAnLZL9>