UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

CURSO DE ESPECIALIZACION DE INGENIERIA DE DATOS

DESARROLLO E IMPLEMENTACION DE UN MODELO DIMENSIONAL PARA EL ANALISIS DE INFORMACION SOBRE ACCIDENTES DE TRANSITO EN LA INDIA EN LOS AÑOS DEL 2007 AL 2020

PRESENTADO POR:

HAROLD JEANCARLOS GUEVARA TORRES

GERARDO JOSE MENDOZA CASTRO

PARA OPTAR AL TITULO DE:

INGENIERO(A) DE SISTEMAS INFORMÁTICOS

CIUDAD UNIVERSITARIA, DICIEMBRE 2023

UNIVERSIDAD DE EL SALVADOR

RECTOR:

M.Sc. JUAN ROSA QUINTANILLA

SECRETARIO GENERAL:

LIC. PEDRO ROSALÍO ESCOBAR CASTANEDA

FACULTAD DE INGENIERÍA Y ARQUITECTURA

DECANO:

ING. LUIS SALVADOR BARRERA MANCÍA

SECRETARIO:

ARQ. RAUL ALEXANDER FABIÁN ORELLANA

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DIRECTOR:

ING. CÉSAR AUGUSTO GONZÁLEZ

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

CURSO DE ESPECIALIZACION DE INGENIERIA DE DATOS

Trabajo de Graduación previo a la opción al Grado de:

INGENIERO(A) DE SISTEMAS INFORMÁTICOS

Título:

DESARROLLO E IMPLEMENTACION DE UN MODELO DIMENSIONAL PARA EL ANALISIS DE INFORMACION SOBRE ACCIDENTES DE TRANSITO EN LA INDIA EN LOS AÑOS DEL 2007 AL 2020

Presentado por:

HAROLD JEANCARLOS GUEVARA TORRES

GERARDO JOSE MENDOZA CASTRO

Trabajo de Graduación Aprobado por:

Docente Asesor:

ING. RENE FABRICIO QUINTANILLA GOMEZ

SAN SALVADOR, DICIEMBRE 2023

Trabajo de Graduación Aprobado por:

Docente Asesor:

ING. RENE FABRICIO QUINTANILLA GOMEZ

Tabla de Contenidos

1) Introducción

En la India, los accidentes de tráfico son una preocupación importante debido al gran número de incidentes y sus consecuencias devastadoras. La recopilación y el análisis de datos precisos y completos son fundamentales para comprender y abordar este problema de manera efectiva. Para lograrlo, se propone la creación de un Data Warehouse (almacén de datos) dedicado a los accidentes de tráfico en la India.

El Data Warehouse para los accidentes de tráfico en la India estaría diseñado para recopilar datos de múltiples fuentes, como informes policiales, registros hospitalarios, compañías de seguros y otros organismos relevantes. Estos datos incluirían información sobre la ubicación de los accidentes, el tipo de vehículos involucrados, las lesiones sufridas y otros factores relacionados con los accidentes.

Una vez que los datos se recopilen, se almacenarán en un formato estructurado y homogéneo en el Data Warehouse. Esto permitirá un fácil acceso y consulta de la información, así como el análisis de tendencias y patrones a lo largo del tiempo.

El análisis de los datos almacenados en el Data Warehouse proporcionaría información valiosa para el desarrollo de políticas y medidas de seguridad vial más efectivas. Por ejemplo, se podrían identificar áreas geográficas con altas tasas de accidentes, tipos de vehículos o conductores con mayor propensión a los accidentes y factores de riesgo específicos. Esta información permitiría la implementación de estrategias preventivas, mejoras en la infraestructura vial y campañas de concienciación dirigidas.

En resumen, la creación de un Data Warehouse para los accidentes de tráfico en la India tiene como objetivo centralizar y analizar datos relacionados con los accidentes de tráfico en el país. Esto proporcionaría una visión más clara de los factores que contribuyen a estos accidentes y permitiría la implementación de medidas más efectivas para mejorar la seguridad vial. Al aprovechar la potencia del análisis de datos, se espera reducir el número de accidentes y salvar vidas en las carreteras de la India.

2) CAPITULO I: Especificación del Proyecto

a) Situación Actual

i. Antecedentes

Los accidentes de tránsito en India son un problema significativo y complejo que tiene múltiples antecedentes. Algunos de los factores y antecedentes clave incluyen:

Crecimiento del Parque Automotor:

El rápido crecimiento económico en India ha llevado a un aumento sustancial en el número de vehículos en las carreteras. El aumento del parque automotor ha contribuido a una mayor congestión vial y a un aumento en la probabilidad de accidentes.

Infraestructura Vial Deficiente:

La calidad de la infraestructura vial en algunas regiones de India es deficiente. Carreteras en mal estado, falta de señalización adecuada y deficiencias en la planificación urbana pueden contribuir a condiciones peligrosas para la conducción.

Condiciones de Tráfico Desafiantes:

En muchas áreas urbanas, el tráfico es caótico y congestionado. La falta de un sistema de transporte público eficiente puede llevar a un mayor uso de vehículos privados, exacerbando los problemas de tráfico y aumentando el riesgo de accidentes.

Comportamiento de Conducción:

El comportamiento de conducción, incluido el exceso de velocidad, el incumplimiento de las normas de tráfico y la conducción imprudente, contribuye significativamente a la incidencia de accidentes. La conciencia y la cultura de seguridad vial son áreas que requieren atención.

Carencia de Educación Vial:

La falta de conciencia y educación vial adecuada es un antecedente importante. Muchos conductores y peatones pueden no estar completamente informados sobre las normas de tráfico y las mejores prácticas de seguridad vial.

Falta de Cumplimiento Normativo:

La falta de aplicación efectiva de las normas de tráfico puede permitir comportamientos peligrosos en las carreteras. La vigilancia y la aplicación de las leyes de tránsito son esenciales para disuadir a conductores irresponsables.

Problemas de Seguridad de Vehículos:

Algunos vehículos pueden no cumplir con estándares de seguridad adecuados, lo que aumenta el riesgo de lesiones graves en caso de accidente. La implementación de normas más estrictas de seguridad vehicular puede ser un factor clave para abordar este problema.

Condiciones Climáticas y Geográficas:

En algunas regiones de India, las condiciones climáticas y geográficas pueden contribuir a la peligrosidad de las carreteras. Por ejemplo, las lluvias monzónicas pueden afectar la visibilidad y la tracción en la carretera.

Crecimiento Urbano Desordenado:

El crecimiento urbano desordenado, con una planificación insuficiente, puede resultar en una falta de infraestructuras viales adecuadas y contribuir a la congestión del tráfico.

Problemas Socioeconómicos:

Factores socioeconómicos, como la falta de acceso a la educación y a oportunidades económicas, pueden contribuir a un comportamiento de conducción arriesgado y a un mayor riesgo de accidentes.

Estos antecedentes subrayan la complejidad del problema de los accidentes de tránsito en India, destacando la necesidad de enfoques integrales que aborden aspectos de infraestructura, educación, aplicación de normas y cultura de seguridad vial.

ii. Descripción del problema

Los accidentes en la india han alcanzado números significativos generando una gran cantidad de datos que pueden ser estudiados para mejorar la situación actual; como grupo hemos decidido crear una serie de Dashboards que nos permita ver el comportamiento de estos datos.

Para ello vemos la necesidad utilizar las herramientas como AWS S3, IAM, AWS Redshift, Talend Open Studio, Power BI, para poder obtener información necesaria que ayude a la toma de decisiones a de los interesados para ello nos enfocaremos en dar respuesta a las siguientes interrogantes.

- El total de accidentes de tránsito durante un periodo definido.
- La cantidad de lesionados por género y edad en los accidentes de tránsito.
- La cantidad de accidentes por ubicación definida.
- El total de accidentes por tipo de vehículo (camioneta, sedan, microbús, etc)
- El total de víctimas por tipo de accidente

iii. Planteamiento del problema

Se realizo el planteamiento del problema en mediante el uso de la caja negra en el cual tendremos los siguientes implicados.

Entradas:

Estas serán los datos obtenidos principalmente mediante Kaggle y también de otras fuentes de datos.

Procesos

En esta etapa se usarán diferentes herramientas en el análisis de los datos, los cuales son y Talend Open Studio para poder realizar la extracción, transformación y carga; AWS Services (IAM, S3, Redshift) para tener una conexión con Talend Open Studio y poder cargar los datos a la nube, finalizando con la herramienta Power BI para transformar, visualizar datos y crear informes de los mismos.

Salidas:

Los Informes Gráficos representará los diferentes reportes que se esperan obtener de la transformación de los datos los cuales se nombran a continuación: Cantidad de accidentes periódicamente, Total de compras por fecha, Total de personas fallecidas por género, Total de personas lesionadas por Genero, edades promedio de accidentes, Porcentaje anuales de vehículos implicados

b) Objetivos

Elaborar reportes analíticos a partir de una solución de data Warehouse que permita de manera concreta identificar la cantidad de accidentes ocurridos en periodos de tiempos específicos contabilizando las víctimas y conductores heridos y fallecidos, así como también la geolocalización de los accidentes.

c) Alcances

Para el proyecto de desarrollo e implementación de un modelo dimensional para el análisis de la información sobre accidentes de tránsito en la india en los años del 2007 al 2020 presentan a continuación los siguientes alcances:

- Brindar un esquema estrella del modelo dimensional que tendrá la potencialidad analítica para poder responder a las necesidades analíticas actuales que originaron el proyecto, así como también está diseñado para poder tener la capacidad de poder responder a futuros requerimientos analíticos o necesidades relacionadas al proceso de negocio de preguntas y respuestas.

- Creación de procesos que extraigan, transformen y carguen la información generada por el sistema transaccional.

- Brindar una solución de Data Warehouse sobre los accidentes de tránsitos en la india para que pueden apoyarse en el proceso de toma de decisiones basadas en sus propios datos.

- Presentación de Dashboards que den solución a las necesidades analíticas que originaron el proyecto, de tal manera que los resultados permitirán ver el auge de los accidentes de tránsito en periodos de tiempo

d) Justificación

La india es uno de los países en los cuales se generan grandes cantidades de accidentes al año, esto genera un impacto fuerte en la economía del país y su población así mismo se pretende determinar cuáles son las zonas geográficas más afectadas por accidentes, en las cual es el género (masculino o femenino) que más se encuentran implicados en ellos también determinar un censo de las personas fallecidas por género y en rangos de edades, así como las lesionadas que logran sobrevivir. Los tipos de vehículos implicados también son una fuerte variable a estudiar para determinar factores de riesgos.

Esta data generada servirá para empresas de seguro para la toma oportuna de decisiones en planes de seguros automovilísticos para ofrecer

e) Cronograma de Actividades

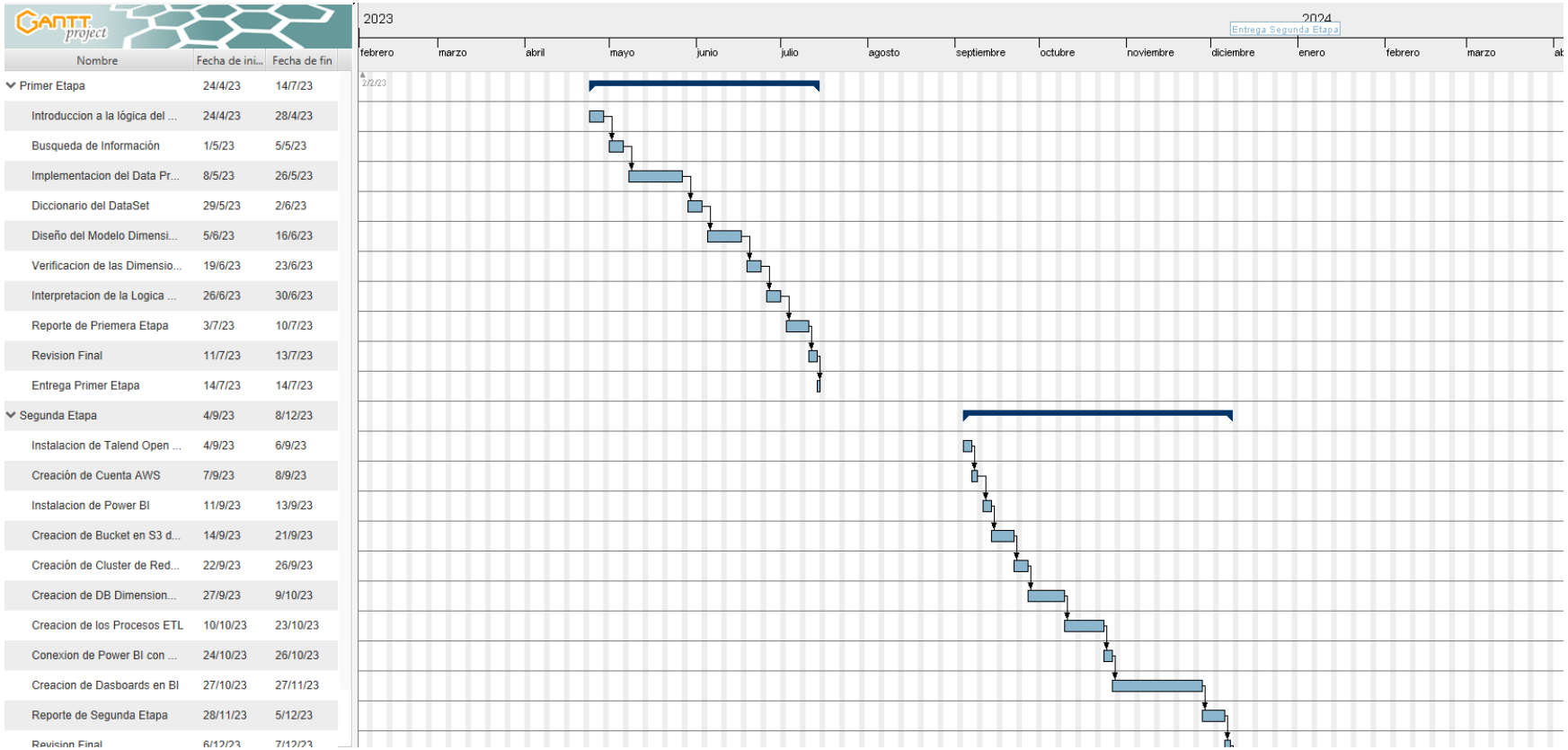


Figura 1. Cronograma de actividades

f) Presupuesto de Desarrollo

Se debe contar con un almacenamiento en la nube, para esta solución se tomará en cuenta Amazon Web Services, del cual se optará por usar los servicios de IAM, Redshift y S3 que nos ofrece esta plataforma. Los precios estimados mensuales y anuales se obtuvieron por medio de AWS Pricing Calculator un servicio con el que cuenta la plataforma, el presupuesto se detalla a continuación:

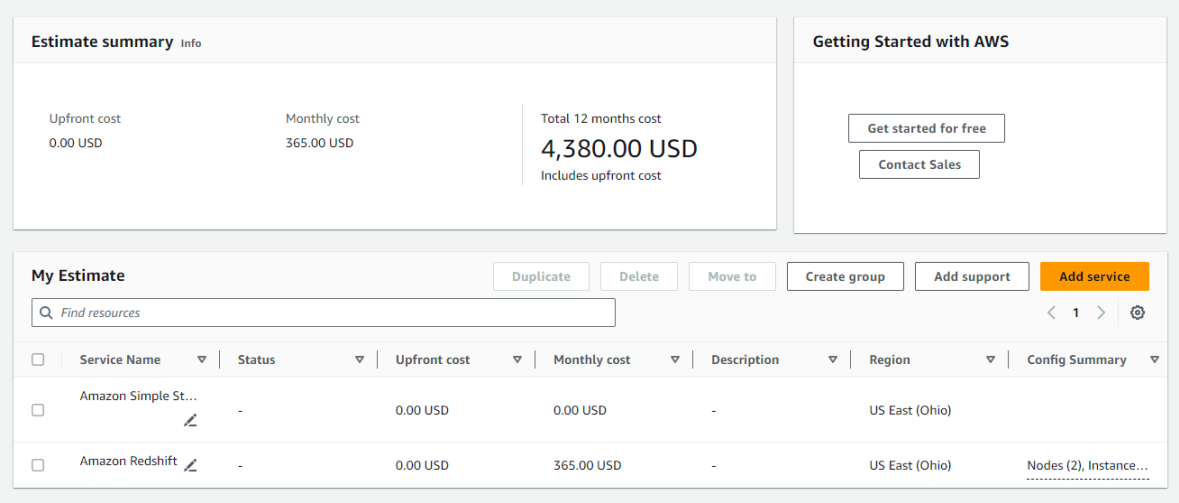


Figura 2. Cálculo de Presupuesto

Otros Costos

Nombre de Costo	Valor	Cantidad	Periodo	Total
Recurso Humano	$ 800	2	3 meses	$ 4800
Equipo Informático	$ 550	2		$ 1100
Gastos de Servicios Básicos	$ 40	1		$ 120
TOTAL	$ 6020			

Tabla 1. Tabla de Costos Totales

g) Marco Teórico

Historia del Análisis de datos.

Una descripción más útil y moderna sugerida es que el "análisis de datos" es una herramienta importante para obtener información comercial y brindar respuestas personalizadas a los clientes. El análisis de datos, a veces abreviado como "análisis", se ha vuelto cada vez más importante para organizaciones de todos los tamaños. La práctica del análisis de datos ha evolucionado y ampliado gradualmente con el tiempo, brindando muchos beneficios.

El uso de análisis por parte de las empresas se remonta al siglo XIX, cuando Frederick Winslow Taylor inició ejercicios de gestión del tiempo. Otro ejemplo es cuando Henry Ford midió la velocidad de las cadenas de montaje. A fines de la década de 1960, Analytics comenzó a recibir más atención a medida que las computadoras se convirtieron en sistemas de apoyo para la toma de decisiones. Con el desarrollo de Big data, almacenes de datos, la nube y una variedad de software y hardware, el análisis de datos ha evolucionado significativamente. El análisis de datos implica la investigación, el descubrimiento y la interpretación de patrones dentro de los datos. Las formas modernas de análisis de datos se han ampliado para incluir:

a) Análisis predictivo
b) Análisis de grandes datos
c) Análisis cognitivo
d) Analítica prescriptiva
e) Analítica descriptiva
f) Gestión de decisiones empresariales
g) Análisis minorista
h) Análisis aumentado
i) analista de la red
j) Análisis de llamadas

Estadística y Computación

El análisis de datos se basa en estadísticas. Se ha supuesto que las estadísticas se utilizaron desde el Antiguo Egipto para construir pirámides. Los gobiernos de todo el mundo han utilizado estadísticas basadas en censos para una variedad de actividades de planificación, incluida la fiscalidad. Una vez que se han recopilado los datos, comienza el objetivo de descubrir información y conocimientos útiles. Por ejemplo, un análisis del crecimiento de la población por departamento y ciudad podría determinar la ubicación de un nuevo hospital.

El desarrollo de las computadoras y la evolución de la tecnología informática ha mejorado drásticamente el proceso de análisis de datos. En 1880, antes de las computadoras, la Oficina del Censo de EE. UU. tardó más de siete años en procesar la información recopilada y completar un informe final. En respuesta, el inventor Herman Hollerith produjo la "máquina tabuladora", que se utilizó en el censo de 1890. La máquina tabuladora podría procesar sistemáticamente los datos registrados en las tarjetas perforadas. Con este dispositivo, el censo de 1890 se terminó en 18 meses.

Bases de datos relacionales y bases de datos no relacionales

Las bases de datos relacionales fueron inventadas por Edgar F. Codden la década de 1970 y se hicieron muy populares en la década de 1980. Las bases de datos relacionales (RDBM), a su vez, permitieron a los usuarios escribir en secuencia (SQL) y recuperar datos de su base de datos.

Las bases de datos relacionales y SQL brindaron la ventaja de poder analizar datos a pedido y todavía se usan ampliamente. Es fácil trabajar con ellos y muy útiles para mantener registros precisos. En el lado negativo, los RDBM generalmente son bastante rígidos y no fueron diseñados para traducir datos no estructurados.

A mediados de la década de 1990, Internet se volvió extremadamente popular, pero las bases de datos relacionales no podían seguir el ritmo. El inmenso flujo de información combinado con la variedad de tipos de datos provenientes de muchas fuentes diferentes, dio lugar a bases de datos no relacionales, también conocidas como NoSQL. Una base de datos NoSQL puede traducir datos usando diferentes idiomas y formatos rápidamente y evita la rigidez de SQL reemplazando su almacenamiento "organizado" con mayor flexibilidad.

El desarrollo de NoSQL fue seguido por cambios en Internet. Larry Page y Sergey Brin 3diseñaron el motor de búsqueda de Google para buscar en un sitio web específico, mientras procesan y analizan Big data en computadoras distribuidas. El motor de búsqueda de Google puede responder en unos segundos con los resultados deseados. Los principales puntos de interés del sistema son su escalabilidad, automatización y alto rendimiento.

Almacenes de datos

A fines de la década de 1980, la cantidad de datos recopilados siguió creciendo significativamente, en parte debido a los menores costos de las unidades de disco duro. Durante este tiempo, la arquitectura de los almacenes de datos se desarrolló para ayudar a transformar los datos provenientes de los sistemas operativos en sistemas de apoyo a la toma de decisiones.

Los almacenes de datos son normalmente parte de la nube o parte del servidor central de una organización. A diferencia de las bases de datos relacionales, un almacén de datos normalmente está optimizado para un tiempo de respuesta rápido a las consultas. En un almacén de datos, los datos a menudo se almacenan mediante una marca de tiempo y los comandos de operación, como eliminar y actualizar, se usan con menos frecuencia. Si todas las transacciones de ventas se almacenaran usando marcas de tiempo, una organización podría usar un almacén de datos para comparar las tendencias de ventas de cada mes.

Inteligencia de negocios

El término inteligencia empresarial (BI) se utilizó por primera vez en 1865 y luego fue adaptado por Howard Dresner en Gartner en 1989, para describir la toma de mejores decisiones comerciales a través de la búsqueda, recopilación y análisis de los datos acumulados guardados por una organización. Usar el término "inteligencia de negocios" como una descripción de la toma de decisiones basada en tecnologías de datos fue novedoso y con visión de futuro. Las grandes empresas primero adoptaron BI en la forma de analizar los datos de los clientes de manera sistemática, como un paso necesario para tomar decisiones comerciales.

Procesamiento de datos

La minería de datos comenzó en la década de 1990 y es el proceso de descubrir patrones dentro de grandes conjuntos de datos. El análisis de datos de formas no tradicionales proporcionó resultados sorprendentes y beneficiosos. El uso de la minería de datos surgió directamente de la evolución de las tecnologías de bases de datos y almacenes de datos. Las nuevas tecnologías permiten a las organizaciones almacenar más datos, sin dejar de analizarlos de forma rápida y eficiente. Como resultado, las empresas comenzaron a predecir las necesidades potenciales de los clientes, basándose en un análisis de sus patrones de compra históricos.

Sin embargo, los datos pueden ser malinterpretados. Alguien en los oficios, habiendo comprado dos pares de jeans azules en línea, probablemente no querrá comprar jeans por otros dos o tres años. Dirigirse a esta persona con anuncios de blue jeans es tanto una pérdida de tiempo como irritante para el cliente potencial.

Big Data

En 2005, Roger Magoulas 4le dio ese nombre a Big data. Estaba describiendo una gran cantidad de datos, que parecían casi imposibles de manejar con las herramientas de Business Intelligence disponibles en ese momento. En el mismo año, Hadoop podía procesar grandes volúmenes de datos. La base de Hadoop se basó en otro marco de software de código abierto llamado Nutch, que luego se fusionó con MapReduce de Google.

Apache Hadoop es un marco de software de código abierto, que puede procesar datos estructurados y no estructurados, transmitidos desde casi todas las fuentes digitales. Esta flexibilidad permite que Hadoop (y sus marcos hermanos de código abierto) procesen Big data. A fines de la década de 2000, surgieron varios proyectos de código abierto, como Apache Spark y Apache Cassandra, para enfrentar este desafío.

Analítica en la Nube

En su forma inicial, la nube era una frase que se usaba para describir el "espacio vacío" entre los usuarios y el proveedor. Luego, en 1997, el profesor de la Universidad de Emory, Ramnath Chellappa, describió la computación en la nube como un nuevo "paradigma informático donde los límites de la computación estarán determinados por la lógica económica, en lugar de los límites técnicos únicamente".

En 1999, Salesforce proporcionó un ejemplo muy temprano de cómo usar la computación en la nube con éxito. Aunque primitivo para los estándares actuales, Salesforce usó el concepto para desarrollar la idea de entregar programas de software a través de Internet. Los programas (o aplicaciones) pueden ser accedidos o descargados por cualquier persona con acceso a Internet. Un gerente de la organización podría comprar software en un método bajo demanda rentable sin salir de la oficina. A medida que las empresas y las organizaciones obtuvieron una mejor comprensión de los servicios y la utilidad de la nube, ganó popularidad.

La nube ha evolucionado significativamente desde 1999, con clientes que “alquilan los servicios”, en lugar de adquirir hardware y software con el mismo propósito. Los proveedores ahora son responsables de la resolución de problemas, las copias de seguridad, la administración, la planificación de la capacidad y el mantenimiento. Y, para varios proyectos empresariales, la nube es simplemente más fácil y eficiente de usar. La nube ahora tiene cantidades significativamente grandes de almacenamiento, disponibilidad para múltiples usuarios simultáneamente y la capacidad de manejar múltiples proyectos.

Análisis predictivo

El análisis predictivo se utiliza para hacer pronósticos sobre tendencias y patrones de comportamiento. El análisis predictivo utiliza varias técnicas tomadas de estadísticas, modelado de datos, minería de datos, inteligencia artificial y aprendizaje automático para analizar datos al hacer predicciones. Los modelos predictivos pueden analizar datos actuales e históricos para comprender a los clientes, los patrones de compra, los problemas de procedimiento y predecir peligros y oportunidades potenciales para una organización.

El análisis predictivo comenzó en la década de 1940, cuando los gobiernos comenzaron a usar las primeras computadoras. Aunque ha existido durante décadas, el análisis predictivo ahora se ha convertido en un concepto cuyo momento ha llegado. Con más y más datos disponibles, las organizaciones han comenzado a usar análisis predictivos para aumentar las ganancias y mejorar su ventaja competitiva. El crecimiento continuo de los datos almacenados, combinado con un interés cada vez mayor en el uso de datos para obtener Business Intelligence, ha promovido el uso de análisis predictivos.

Análisis cognitivo

La mayoría de las organizaciones manejan datos no estructurados. Dar sentido a estos datos no estructurados no es algo que los humanos puedan hacer fácilmente. El análisis cognitivo combina una variedad de aplicaciones para proporcionar contexto y respuestas. Las organizaciones pueden recopilar datos de varias fuentes diferentes, y el análisis cognitivo puede examinar los datos no estructurados en profundidad, ofreciendo a los responsables de la toma de decisiones una mejor comprensión de sus procesos internos, las preferencias de los clientes y la lealtad de los mismos.

Análisis aumentado

El análisis aumentado proporciona Business Intelligence (y conocimientos) automatizados mediante el uso de procesamiento de lenguaje natural y aprendizaje automático. "Automatiza" la preparación de datos y permite compartir datos. El análisis aumentado proporciona resultados claros y acceso a herramientas sofisticadas, lo que permite a los investigadores y gerentes tomar decisiones diarias con un alto grado de confianza. Permite a los responsables de la toma de decisiones obtener información y actuar con rapidez y confianza.

En última instancia, análisis aumentado intenta reducir el trabajo de los científicos de datos mediante la automatización de los pasos utilizados, para obtener información e inteligencia comercial. Un motor de análisis aumentado procesará automáticamente los datos de una organización, los limpiará, los analizará y luego producirá información que conducirá a instrucciones para ejecutivos o vendedores.

Análisis de cartera

El análisis de cartera suele ser utilizado por una agencia de préstamos o un banco, y es una colección de cuentas con valores y riesgos variables. Las cuentas en cartera pueden incluir información sobre el estatus social de sus clientes (pobre, clase media, rico), su ubicación geográfica y muchos otros factores. El análisis de cartera permite al prestamista equilibrar los rendimientos de un préstamo con el riesgo de incumplimiento. El riesgo del préstamo está

determinado por factores como los ingresos, el éxito de préstamos anteriores y las declaraciones de quiebra.

Analítica de recursos humanos

Originalmente llamado "análisis de personas", el análisis de recursos humanos son datos de comportamiento que se utilizan para comprender cómo trabajan las personas y cómo cambian la forma en que se administran las organizaciones. El análisis de recursos humanos también se ha denominado análisis de la fuerza laboral, análisis de talento, información de talento, información de personas, información de colegas y análisis de capital humano. El análisis de recursos humanos se utiliza para ayudar a las empresas a administrar sus recursos humanos y es una herramienta estratégica para analizar y pronosticar tendencias en los mercados laborales.

Análisis del viaje del cliente

El viaje del cliente se ocupa de la experiencia holística por la que pasan los clientes al interactuar con una organización o marca. En lugar de centrarse en una parte de la experiencia, el viaje del cliente registra la experiencia completa de un cliente.
El análisis del viaje del cliente examina la información registrada y proporciona información sobre las experiencias del cliente (a menudo en tiempo real). Ayuda a comprender al cliente e influye en cómo las empresas diseñan la experiencia del cliente. El análisis del viaje del cliente admite un método sistemático para evaluar y monitorear el viaje del cliente y mejorar el proceso. Desarrollar y brindar una experiencia óptima al cliente es el objetivo final.

Data Warehouse.

Actualmente es fácil perderse cuando se lidia con datos, existen muchos tipos de datos, cada tipo tiene sus propias peculiaridades e idiosincrasias. En las organizaciones normalmente cada departamento maneja los datos a su propia manera, tienen sus propias aplicaciones, etc. Esto hace difícil que los datos de todos los departamentos se complementen entre sí. Este era el problema al que se enfrentaron muchas organizaciones anteriormente, se dieron cuenta que tener datos, no era lo mismo a tener datos creíbles, se tuvo una discusión acerca de que es "Integridad de datos", y fue precisamente eso fue lo que hizo que el data Warehouse naciera.
Según Kimbal6, un data Warehouse se puede definir una copia de los datos transaccionales, específicamente estructurados para consultas y análisis. Es decir, es el sistema que extrae, transforma y consolida los datos de los sistemas fuentes en un repositorio de datos dimensional.

Características de un Data Warehouse según Bill Immon

- Orientado a temas: los datos están organizados por temas para facilitar el entendimiento por parte de los usuarios, de forma que todos los datos relativos a un mismo elemento de la vida real queden unidos entre sí. Por ejemplo, todos los datos de un cliente pueden estar consolidados en una misma tabla, todos los datos de los productos en otra, y así sucesivamente.

- Integrado: los datos se deben integrar en una estructura consistente, debiendo eliminarse las inconsistencias existentes entre los diversos sistemas operacionales. La

información se estructura en diversos niveles de detalle para adecuarse a las necesidades de consulta de los usuarios. Algunas de las inconsistencias más comunes que nos solemos encontrar son: en nomenclatura, en unidades de medida, en formatos de fechas, múltiples tablas con información similar.

- Histórico (variante en el tiempo): los datos, que pueden ir variando a lo largo del tiempo, deben quedar reflejados de forma que al ser consultados reflejen estos cambios y no se altere la realidad que había en el momento en que se almacenaron, evitando así la problemática que ocurre en los sistemas operacionales, que reflejan solamente el estado de la actividad de negocio presente. Un Data Warehouse debe almacenar los diferentes valores que toma una variable a lo largo del tiempo. Por ejemplo, si un cliente ha vivido en tres ciudades diferentes, debe almacenar el periodo que vivió en cada una de ellas y asociar los hechos (ventas, devoluciones, incidencias, etc.) que se produjeron en cada momento a la ciudad en la que vivía cuando se produjeron, y no asociar todos los hechos históricos a la ciudad en la que vive actualmente.

- No volátil: la información de un Data Warehouse, una vez introducida, debe ser de sólo lectura, nunca se modifica ni se elimina, y ha de ser permanente y mantenerse para futuras consultas. Por ejemplo, si en el origen se modifica la cantidad de un producto que entra en el almacén, en el Data Warehouse no podemos hacer directamente una actualización sobre ese registro sin dejar ni el más mínimo rastro de que hubo antes otro valor.

Esquema en Estrella

A la hora de modelar el Data Warehouse, hay que decidir cuál es el esquema más apropiado para obtener los resultados que queremos conseguir. Habitualmente, y salvo excepciones, se suele modelar la base de datos utilizando el esquema en estrella (star schema), en el que hay una única tabla central, la tabla de hechos, que contiene todas las medidas y una tabla adicional por cada una de las perspectivas desde las que queremos analizar dicha información, es decir por cada una de las dimensiones.

Modelado Dimensional

El Modelado Dimensional es utilizado hoy en día en la mayoría de las soluciones de BI. Es una mezcla correcta de normalización y desnormalización, comúnmente llamada Normalización Dimensional. Se utiliza tanto para el diseño de Data Marts como de Data Warehouses.
Básicamente hay dos tipos de tablas:

- Tablas de Dimensión (Dimension Tables)
- Tablas de Hechos (Fact Tables)

Tabla de hechos

Los Hechos están compuestos por los detalles del proceso de negocio a analizar, contienen datos numéricos y medidas (métricas) de Negocio a analizar. Contienen también elementos

(claves externas) para contextualizar dichas medidas, como por ejemplo el producto, la fecha, el cliente, la cuenta contable, etc.

Componentes de una tabla de hechos

- Clave principal: identifica de forma única cada fila. Al igual que en los sistemas transaccionales toda tabla debe tener una clave principal, en una tabla de hechos puede tenerla o no, y esto tiene sus pros y sus contras, pero ambas posturas son defendibles.

- Claves externas (Foreign Keys): apuntan hacia las claves principales (claves subrogadas) de cada una de las dimensiones que tienen relación con dicha tabla de hechos.

- Medidas (Measures): representan columnas que contienen datos cuantificables, numéricos, que se pueden agregar. Por ejemplo, cantidad, importe, precio, margen, número de operaciones, etc.

- Metadatos y linaje: nos permite obtener información adicional sobre la fila, como, por ejemplo, que día se incorporó al Data Warehouse, de qué origen proviene (si tenemos varias fuentes), etc. No es necesario para el usuario de negocio, pero es interesante analizar en cada tabla de hechos qué nos aporta y si merece pena introducir algunas columnas de este tipo.

Dimensiones

Una dimensión contiene una serie de atributos o características, por las cuales podemos agrupar, rebanar o filtrar la información. A veces estos atributos están organizados en jerarquías que permiten analizar los datos de forma agrupada, dicha agrupación se realiza mediante relaciones uno a muchos (1:N). Por ejemplo, en una dimensión Fecha es fácil que encontremos una jerarquía formada por los atributos Año, Mes y Día, otra por Año, Semana y Día; en una dimensión Producto podemos encontrarnos una jerarquía formada por los atributos Categoría, Subcategoría y Producto

Tipos de claves de una dimensión

- Una Clave subrogada (subrogate key): es un identificador único que es asignado a cada fila de la tabla de dimensiones, en definitiva, será su clave principal. Esta clave no tiene ningún sentido a nivel de negocio, pero la necesitamos para identificar de forma única cada una de las filas. Son siempre de tipo numérico, y habitualmente también son auto-incrementales. En el caso de SQL Server recomendamos que sean de tipo INT con la propiedad identity activada (es una recomendación genérica, a la que siempre habrá excepciones).

- Una Clave natural: es una clave que actúa como primary key en nuestro origen de datos, y es con la que el usuario está familiarizado, pero no puede ser clave principal en nuestra tabla de dimensiones porque se podrían producir duplicidades, como veremos más adelante al explicar el concepto de Slowly Changing Dimensions.

Infraestructura de data Warehouse

La arquitectura de un Data Warehouse varía según el autor, a continuación, presentamos la arquitectura de Immon, la cual es una de las más aceptables en el área de desarrollo de modelos dimensionales.

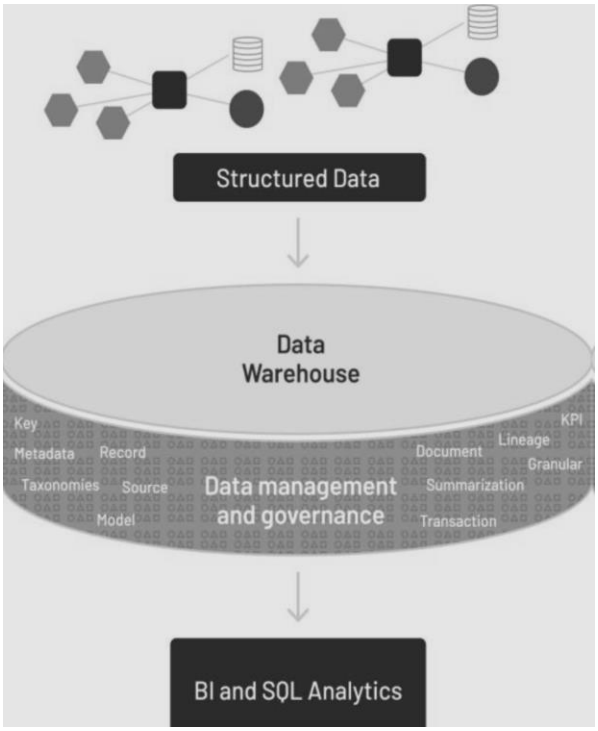


Figura 3. Infraestructura de un Data Werehouse

La infraestructura analítica incluye:

- Metadatos: una guía sobre qué donde se ubicaron que datos.
- Modelo de datos: una abstracción de los datos encontrados en el almacén de datos.
- Linaje de datos: la historia de los orígenes y transformaciones de datos encontrados en los depósitos de datos.
- Resumen: una descripción del algoritmo para trabajar en la creación de los datos en el almacén de datos.
- KPI: ¿dónde están los indicadores clave de rendimiento.
- ETL: tecnología que permitió que los datos de las aplicaciones se transformaren automáticamente en datos corporativos.

La limitación del Data Warehouse se vuelve evidente con el incremento de la variedad de datos (texto, IoT, imágenes, audio, video, etc.) de la organización. En adición, con el surgimiento del Machine learning (ML) e Inteligencia artificial, los nuevos algoritmos requieren acceso directo a los datos a través de lenguajes que no son SQL lo cual dificulta la extracción.

Big Data y Cloud computing.

Origenes de la Big Data

Se han realizado varios estudios sobre las perspectivas históricas y de desarrollo del análisis de big data. Gil Press7 proporciona una breve historia del big data que comenzó en 1944 y cubre los 68 años de historia del desarrollo del big data desde 1944 hasta 2012, ilustrando 32 eventos de big data en la historia moderna de la ciencia de datos. Como señala Press en su artículo, la línea entre el crecimiento de los datos y el big data es borrosa. Muchas veces se hace referencia a la tasa de crecimiento de los datos como la "explosión de la información"; Aunque "datos" e "información" se utilizan a menudo indistintamente, los dos términos tienen significados diferentes. La investigación realizada por la prensa cubre eventos hasta 2013, pero muchos eventos cubren big data y ciencia de datos, por lo que el término "ciencia de datos" puede considerarse un significado complementario al análisis de big data. (Oficina de Desarrollo Empresarial).

Contrariamente a las investigaciones del editor, Frank Ohlhorst identifica el origen de los grandes datos en el décimo censo de Estados Unidos de 1880. El problema que surgió en el siglo XIX fue la estadística, que consistía esencialmente en encuestar y registrar a los 50 millones de ciudadanos de América del Norte. Aunque big data puede contener algunos elementos de computación estadística, los dos términos tienen hoy interpretaciones diferentes. Al igual que Frank Ohlhorst, muchos coinciden en que este ciclo es el origen del big data, argumentando que cuando un conjunto de datos es tan grande y complejo que excede las capacidades tradicionales de procesamiento y gestión, ese conjunto de datos puede considerarse big data. datos.

¿Qué es Big Data?

Big data se refiere a la gran y variada cantidad de información que crece a un ritmo cada vez mayor. Esto incluye la cantidad de información, la velocidad a la que se crea y recopila y la variedad o rango de puntos de datos cubiertos (conocidos como las “tres V” del big data). Los big data a menudo provienen de la minería de datos y se presentan en múltiples formatos.

Características claves de Big Data

- Big data es una gran cantidad de información diversa que llega en volúmenes crecientes y con una velocidad cada vez mayor.
- Los macrodatos pueden ser estructurados (generalmente digitales y fáciles de formatear y almacenar) o no estructurados (más libres y menos mensurables).
- Casi todos los departamentos de una empresa pueden utilizar los resultados del análisis de big data, pero gestionar su caos y ruido puede resultar problemático.
- Los big data pueden recopilarse a partir de comentarios compartidos públicamente en redes sociales y sitios web, así como voluntariamente desde dispositivos y aplicaciones electrónicos personales, a través de cuestionarios, compras de productos y registros electrónicos.

- Los macrodatos a menudo se almacenan en bases de datos informáticas y se analizan mediante software diseñado específicamente para procesar conjuntos de datos grandes y complejos.

Ventajas y Desventajas de Big Data

El aumento de los datos disponibles crea oportunidades y desafíos. En general, más datos sobre los clientes (y clientes potenciales) permitirían a las empresas adaptar mejor los productos y los esfuerzos de marketing para generar los niveles más altos de satisfacción y repetición de negocios. Las empresas que recopilan grandes cantidades de datos tienen la oportunidad de realizar análisis más profundos y ricos para beneficiar a todas las partes interesadas.

Si bien un mejor análisis es positivo, el big data también genera gastos generales y ruido, lo que reduce su utilidad. Las empresas deben procesar grandes cantidades de datos y determinar qué datos representan una señal y cuáles representan ruido. Decidir qué hace que los datos sean significativos se convierte en un factor clave.

Además, debido a la naturaleza y formato del material, es posible que se requiera un procesamiento especial antes de poder insertarlo. Los datos estructurados que constan de valores numéricos se pueden almacenar y clasificar fácilmente. Los datos no estructurados, como mensajes de correo electrónico, vídeos y archivos de texto, pueden requerir métodos más sofisticados.

Cloud computing

Cloud computing o computación en la nube son servicios de recursos computacionales distribuidos a través de la red. Dichos recursos computacionales son presentados como uno o muchos recursos unificados, esto según sea el acuerdo del consumidor con el proveedor de servicios.

Modelo de servicio ofrecidos en Cloud Computing

- Software como servicio (SaaS):

En este modelo, se accede a las aplicaciones según sea necesario y el consumidor no administra ni controla la infraestructura, los servidores, los sistemas operativos o el almacenamiento de la nube. Un ejemplo de estos servicios es el correo electrónico.

- Plataforma como servicio (PaaS):

En este modelo, los consumidores implementan sus propias aplicaciones; estas aplicaciones deben desarrollarse utilizando lenguajes y herramientas de programación soportados por el proveedor. El consumidor no gestiona ni controla la infraestructura de la nube, los servidores, los sistemas operativos o el almacenamiento, pero tiene control sobre las aplicaciones instaladas y puede controlar la configuración del entorno de alojamiento.

- Infraestructura como servicio (IaaS):

En este modelo, los consumidores pueden elegir procesamiento, almacenamiento y otros recursos informáticos centrales para que puedan implementar y ejecutar cualquier software, que puede incluir sistemas operativos y aplicaciones. Los consumidores no gestionan ni controlan la infraestructura de la nube; Puede administrar el sistema operativo, el

almacenamiento, las aplicaciones implementadas y tener control limitado sobre los elementos de la red.

Modelos de implementación de Cloud Computing

Independientemente del modelo de servicio utilizado (SaaS, PaaS, IaaS), existen cuatro formas principales de implementar servicios de computación en la nube:

- Nube pública:

Los proveedores de servicios implementan servicios en su propia infraestructura y los ponen a disposición del público. Los principales desafíos de este modelo están relacionados con la seguridad de la información y la calidad del servicio.

- Nube privada:

La infraestructura de la nube es administrada por una organización para brindar servicios de TI a sus usuarios internos. Como ventaja, el centro de datos se vuelve más ágil y flexible y se consigue una mejor gestión de los recursos, pero como desventaja, se pierde la escalabilidad de la nube al estar limitada por los recursos físicos disponibles.

- Nube híbrida:

Este modelo complementa la nube privada con servicios de nube pública para lograr las ventajas de ambos modelos.

Data Lake

Un data lake es un repositorio digital diseñado para almacenar grandes cantidades de datos en su formato original, es decir, su estructura no ha sido modificada. Las fuentes de información en un data lake son muy diversas, incluyendo sistemas transaccionales, sistemas de gestión, logs, correos electrónicos, productos electrónicos, Internet de las cosas, etc.

Los data lake se pueden utilizar para una variedad de propósitos; a continuación, se ofrece una descripción general de los usos más comunes:

- Ingerir datos multiestructurados, semiestructurados y no estructurados.
- Una plataforma ETL que prepara y crea archivos de configuración para sistemas de almacenamiento para que las organizaciones no se vean obligadas a expandir su almacén de datos existente.
- Más que solo almacenamiento de datos, ya que los lagos de datos le permiten almacenar datos que un almacén de datos tradicional no puede manejar fácilmente.
- Archivado de datos y almacenamiento histórico en toda la organización.
- Almacenar información generada por múltiples dispositivos en Internet de las Cosas (IOT) que luego será analizada mediante Machine Learning (ML), etc.

Dado que un lago de datos puede dedicarse a múltiples propósitos, la distribución interna del catálogo dependerá de los propósitos previstos.

Los data lake pueden almacenar datos estructurados, semiestructurados y no estructurados.

- **Datos estructurados:** la mayoría de las organizaciones crean datos estructurados como parte de las transacciones diarias que se escriben en una base de datos SQL. Cuando se ejecuta una transacción, una propiedad importante de los datos es que cada nuevo dato tiene una estructura similar a la de los datos originales. En el pasado.

- **Datos semiestructurados:** Estos datos no tienen un esquema definido, algunos tipos de datos están en formato XLM, JSON, TEXT y estos datos se almacenan en bases de datos NoSQL.

- **Datos no estructurados:** son datos que no siguen ciertas reglas y cada formato de datos requiere un nivel diferente de complejidad para analizar en comparación con los tipos de datos anteriores. Entre algunos datos no estructurados tenemos: imágenes, audio y video.

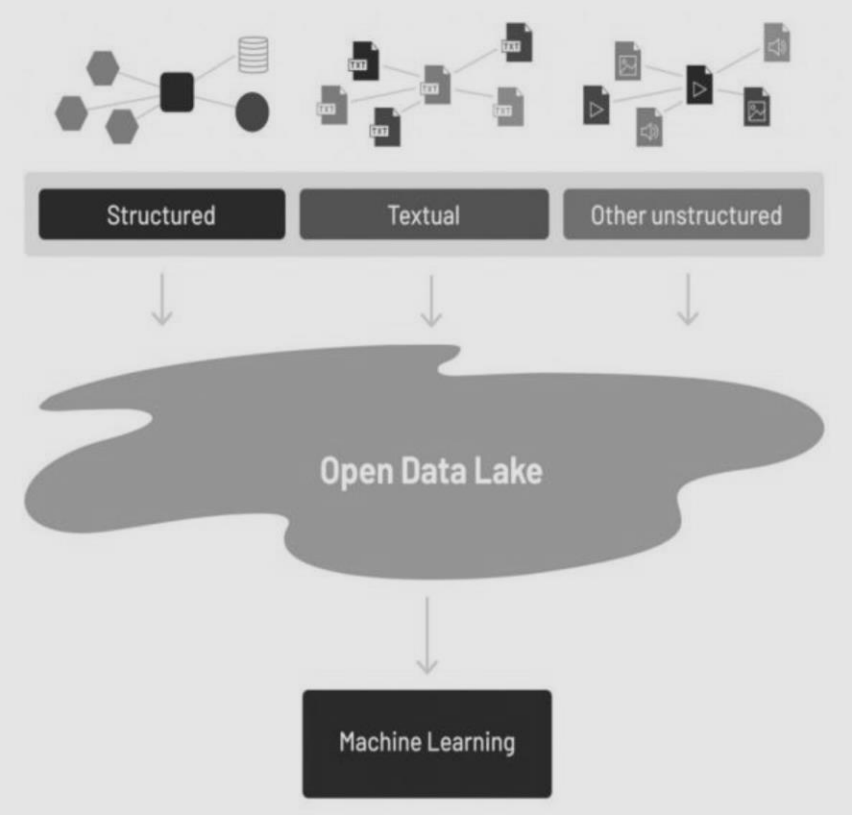


Figura 4. Data Lake

Data Lakehouse

Data Lake House es una nueva arquitectura de gestión de datos que combina los mejores elementos de los lagos de datos y los almacenes de datos. Aprovecha la flexibilidad, el bajo costo y la escalabilidad de un lago de datos y la gobernanza de datos utilizando principios ACID para

el almacenamiento de datos para permitir procesos de BI y aprendizaje automático para todos los datos.

Data Lakehouse le permite aprovechar las capacidades de DW para administrar datos estructurados y administrar datos directamente en el almacenamiento de bajo costo de los lagos de datos. Centralizar estas capacidades en un sistema significa que los equipos de datos pueden moverse rápidamente y consumir datos sin tener que acceder a múltiples sistemas. Esto garantiza que los datos sean los más completos, actualizados y utilizables para proyectos de ciencia de datos, aprendizaje automático y análisis de negocios.
Al igual que un lago de datos, una casa de lago de datos admite la entrada de datos estructurados, semiestructurados y no estructurados. Y le permite administrar datos o administrar con datos ya procesados para luego usarlos para diversas aplicaciones como BI, ML, ciencia de datos, etc.

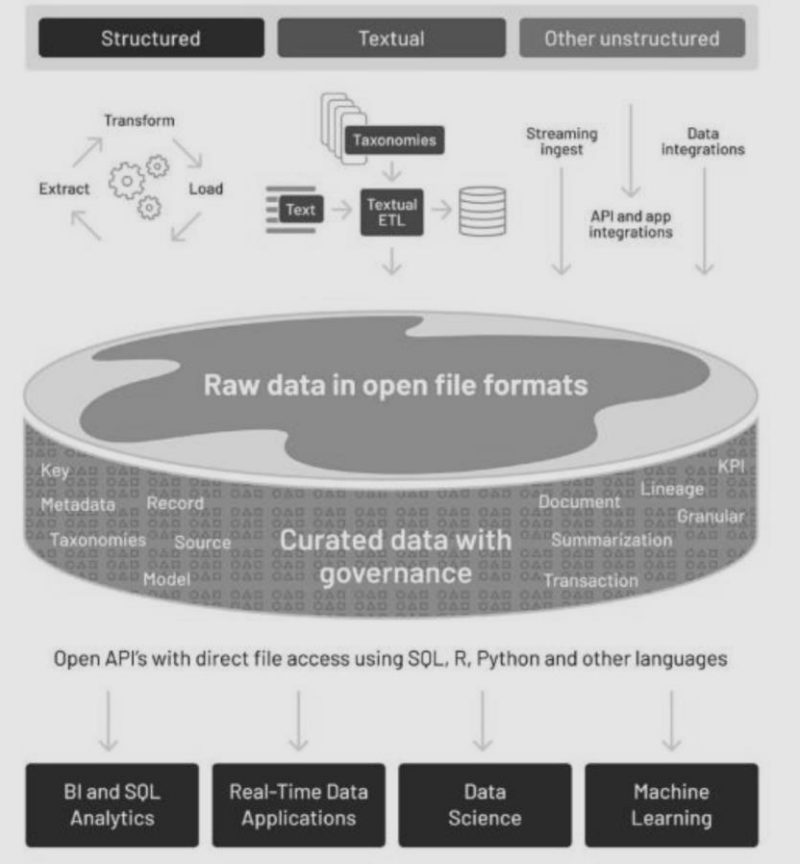


Figura 5. Data LakeHouse

Comparativa entre un Data Warehouse, Data Lake y Data Lakehouse

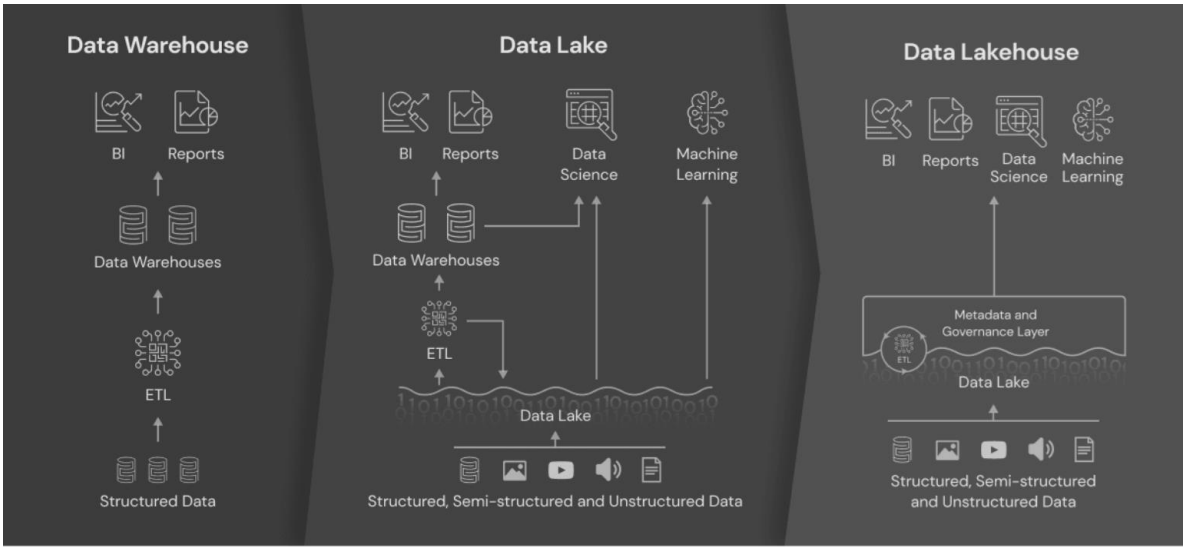


Figura 6. Comparativa entre un Data Warehouse, Data Lake y Data Lakehouse

Descripción de sus elementos.

Planeación del Proyecto

Durante este proceso se determinó el propósito del proyecto DW/BI, los objetivos y alcances específicos, los principales riesgos y el enfoque inicial para satisfacer las necesidades de información. En el plan y la visión del proyecto de Kimball, un proyecto se refiere a una iteración simple de KLC (Ciclo de vida de Kimball) desde el inicio hasta la implementación. Esta tarea incluye las siguientes actividades típicas en el plan del proyecto:

1. Definir el alcance (comprender las necesidades del negocio).
2. Definir tareas
3. Organizar tareas
4. Planificar el uso de los recursos. 5. Asignar cargas de trabajo a recursos
5. Preparar los documentos finales del proyecto.
6. Proyecto.

Además, en este apartado definimos cómo realizar la administración o gestión de estas subfases, que son un proyecto completo en sí mismas, con los siguientes pasos:

1. Monitorear el estado de los procesos y actividades.
2. Seguimiento de problemas
3. Desarrollar un plan de comunicación integral para la empresa y el campo de TI.

Definición de requerimiento del negocio

Antes de comenzar cualquier trabajo de diseño con modelos dimensionales, debe comprender los requisitos comerciales y la realidad de sus fuentes de datos existentes. Es necesario obtener los requisitos de la forma más clara reuniéndose con representantes de la empresa. Hay cuatro decisiones de diseño importantes que se deben tomar con los representantes comerciales. Estas decisiones son:

1. Elección del proceso empresarial
2. tamaño de partícula
3. Identificación del tamaño
4. Identificación métrica.

Las respuestas a estas preguntas se determinarán en función de las necesidades de la empresa y la realidad de los datos disponibles. Una vez que se definen los procesos de negocio, la granularidad, las dimensiones y las métricas requeridas, el equipo de desarrollo puede comenzar a construir el modelo dimensional y su correspondiente implementación.

Diseño de la arquitectura técnica

La definición de requisitos comerciales responde a la pregunta "¿Qué vamos a hacer?" y el diseño arquitectónico responde a la pregunta "¿Cómo queremos lograr esto?"

La arquitectura técnica es el plan maestro para que el almacén de datos esté listo cuando se implemente. Describe el flujo de datos desde los sistemas de fuentes de información a través de la conversión y el almacenamiento de datos hasta los tomadores de decisiones.

Este paso también identifica las herramientas, técnicas, utilidades y plataformas necesarias para hacer fluir datos a través del DW.

Normalmente, una arquitectura DW consta de cuatro servidores: un servidor ETL, un servidor de base de datos, un servidor OLAP y un servidor de informes.

Selección del producto e instalación

Parece una lista de compras y se utiliza para seleccionar productos que encajan en el sistema de planificación. Estas seis tareas de selección de productos relacionadas con DW/BI son muy similares a cualquier selección de tecnología.

Comprender el proceso de compras corporativas.

El primer paso antes de seleccionar nuevos productos es entender el hardware interno y procesos de compra de software.

Desarrollar una matriz de evaluación de productos.

Usando el plan de arquitectura como punto de partida, una evaluación basada en una hoja de cálculo se debe desarrollar una matriz que identifique los criterios de evaluación, junto con la ponderación factores para indicar importancia; cuanto más específicos sean los criterios, mejor.

Si los criterios son demasiado vagos o genéricos, todos los proveedores dirán que pueden satisfacer sus necesidades.

Realizar estudios de mercado.

Para ser un comprador informado al elegir un producto, es necesario realizar una investigación de mercado para comprender mejor a los actores y sus productos. La solicitud de propuesta (RFP) es una herramienta clásica de evaluación de productos.

Aunque algunas organizaciones no tienen otra opción, usted debe evitar utilizar este método si es posible. Crear una propuesta y evaluar las respuestas puede consumir mucho tiempo de su equipo.

Al mismo tiempo, los proveedores de servicios están motivados para responder las preguntas de la manera más positiva, por lo que evaluar las respuestas suele ser más bien un concurso de belleza. Después de todo, es posible que el valor de la tarifa no valga la pena.

Evaluar una lista corta de opciones.

A pesar de la gran cantidad de productos en el mercado, a menudo hay sólo unos pocos proveedores que pueden cumplir con los requisitos técnicos y funcionales. Al comparar las puntuaciones preliminares en la matriz de puntuación, puede centrarse en una lista limitada de proveedores y descalificar al resto. Después de tratar con un número limitado de proveedores, puede iniciar una evaluación detallada.

Si está evaluando herramientas de BI, debe involucrar a representantes de la empresa en el proceso. Como evaluador, usted debe impulsar el proceso, no dejar que las ventas lo impulsen, compartiendo información crítica del plan de arquitectura para mantener la reunión centrada en sus necesidades, no en el producto.

Asegúrese de hablar con las referencias de su proveedor, tanto las que figuran oficialmente como las que obtiene de su red informal.

Si es necesario, realice un prototipo.

Después de una evaluación detallada, a veces surge un claro ganador basado en la experiencia o las relaciones pasadas del equipo. En otros casos, los clientes potenciales surgen de compromisos comerciales existentes, como licencias de sitios o compras de hardware heredado. De cualquier manera, puede omitir el paso de creación del prototipo (y la inversión asociada de tiempo y dinero) si un candidato es el ganador.

Si ningún proveedor es un claro ganador, prototípelo con no más de dos productos. Nuevamente, asuma la responsabilidad del proceso desarrollando estudios de casos de negocios limitados pero realistas.

Seleccionar producto, instalar en prueba y negociar

El producto ahora está seleccionado. En lugar de firmarlo inmediatamente en la línea punteada, pero mantenga sus habilidades de conversación haciendo pasivos personales con un vendedor.

En lugar de decirle a su vendedor que no tiene existencias, inicie una prueba que le brinde la oportunidad de utilizar el producto en su entorno. Instalar, capacitar y comenzar a utilizar el producto requiere mucho esfuerzo, por lo que solo debe trabajar con proveedores a los que tenga plena intención de comprar; Las pruebas no deberían ser como cualquier otro ejercicio de patear neumáticos.

Una vez finalizado el período de prueba, tienes la opción de negociar una compra que es para beneficio de todas las partes involucradas.

Desarrollo de data profiling

Una vez que se obtienen los requisitos de la empresa, se deben revisar las fuentes de datos disponibles que respaldan los requisitos. La mejor manera de lograrlo es mediante el análisis de datos, que ayudará a desarrollar una comprensión más profunda de la estructura de las fuentes de datos, su contenido, las relaciones y las reglas que subyacen a los datos. Es necesario comprobar si el material existe y está en condiciones de ser utilizado. El análisis de datos puede ser tan simple como escribir algunas consultas SQL o tan complejo como usar herramientas especiales.

Los resultados del análisis de datos se registran, generalmente mediante la creación de una lista de elementos de datos que cumplen con características aceptables que se utilizarán más adelante en el proceso de ETL

Modelado dimensional

Características del modelado dimensional.

El modelado dimensional es una técnica de diseño lógico, esta técnica nos proveerá de una estructura de los datos que tengan las siguientes características:

- Un alto rendimiento en la consulta de datos
- Intuitivos para los usuarios del negocio.

Los modelos dimensionales son llamados modelos o esquemas estrellas, estos modelos son guardados en estructuras OLAP conocidas como Cubos OLAP.

Proceso iterativo para crear un modelo dimensional:

1. Seleccione un proceso de negocio:

Este proceso depende del análisis de las necesidades del negocio.

2. Establezca el nivel de granularidad:

El nivel de detalle depende de qué tan detallado quieras llegar. Se recomienda buscar el nivel de detalle más profundo que permitan los datos, dado lo que se requiere en base al análisis de las necesidades del negocio.

3. Elige una talla:

Las dimensiones suelen tener atributos literales que nos darán contexto en el nivel de granularidad elegido. Para identificar dimensiones, se deben analizar sus propiedades para encontrar propiedades candidatas para nombres de informes, cubos o cualquier tipo de visualización unidimensional o multidimensional.

4. Determine las medidas y la tabla de hechos:

Las medidas o indicadores son los valores que se dan en un proceso de negocio. Las medidas o indicadores son características por las cuales se desea analizar, resumir o agrupar datos. Estas medidas o indicadores se almacenan en tablas que se relacionan con las dimensiones que brindan contexto para la medición.

Dimensiones

Las tablas de dimensiones son un complemento completo de las tablas de hechos. Las tablas de dimensiones contienen contextos textuales relacionados con eventos de medición de procesos de negocio.

Las dimensiones intentan describir el "qué, qué, dónde, cuándo, cómo y por qué" asociado con un evento. Las tablas de dimensiones suelen tener muchas columnas o atributos. No es inusual que las tablas de dimensiones tengan entre 50 y 100 atributos, aunque, por supuesto, algunas tablas de dimensiones sólo tienen unos pocos atributos.

Las tablas de dimensiones suelen tener menos filas que las tablas de hechos, pero pueden ser anchas y contener muchas columnas de texto grandes. Cada dimensión está definida por una única clave primaria. Los atributos de dimensión sirven como fuente principal de restricciones de consulta, agrupaciones y etiquetas de informes. Los atributos se identifican mediante palabras separadas en una consulta o solicitud de informe.

Los atributos de la tabla de dimensiones juegan un papel crucial en los sistemas DW/BI. por culpa de ellos
Los atributos de dimensión son la fuente de casi todas las restricciones y etiquetas de informes, y son esenciales para que un sistema DW/BI sea utilizable y comprensible. Los atributos deben construirse como datos textuales que reflejen completamente el contexto de la presentación, y debemos reducir el uso de código en las tablas de dimensiones, reemplazándolos con código mucho más detallado o textual.

Las tablas de dimensiones a menudo representan relaciones jerárquicas. Por ejemplo, los productos se dividen en marcas y luego en categorías. Para cada fila de una categoría de producto, mantenga la marca y la descripción de categoría adecuadas.

En lugar de encontrar una tercera forma normal, generalmente está muy desnuda, tanta relación con las relaciones con una tabla de tamaño único. Debido a que la tabla de tamaño suele ser menor que una tabla con hechos o hechos, en realidad mejora la eficiencia de la estandarización o el almacenamiento de copos de nieve. De hecho, no afecta el tamaño total de los principios básicos. Casi siempre tenemos que sacrificar el espacio de la mesa dimensional por simplicidad y accesibilidad.

Product Key	Product Description	Brand Name	Category Name
1	PowerAll 20 oz	PowerClean	All Purpose Cleaner
2	PowerAll 32 oz	PowerClean	All Purpose Cleaner
3	PowerAll 48 oz	PowerClean	All Purpose Cleaner
4	PowerAll 64 oz	PowerClean	All Purpose Cleaner
5	ZipAll 20 oz	Zippy	All Purpose Cleaner
6	ZipAll 32 oz	Zippy	All Purpose Cleaner
7	ZipAll 48 oz	Zippy	All Purpose Cleaner
8	Shiny 20 oz	Clean Fast	Glass Cleaner
9	Shiny 32 oz	Clean Fast	Glass Cleaner
10	ZipGlass 20 oz	Zippy	Glass Cleaner
11	ZipGlass 32 oz	Zippy	Glass Cleaner

Figura 7. Ejemplo de Dimensiones

La mayoría de las veces, se implementan dimensiones de tipo consistentes asociadas con varias tablas de hechos, es decir, proporcionan contexto no solo para un proceso de negocios, sino para varios procesos de negocios, lo que permite optimizar el proceso de modelado, optimizar el almacenamiento, lo que facilita la creación de modelos, por ejemplo, empresas con gran experiencia en el desarrollo de soluciones big data que han implementado una gran cantidad de modelos, lo que les permite obtener una gran cantidad de elementos del modelo cuando se lanzan nuevos desarrollos. Los bucles de dimensiones comienzan con el modelado de dimensiones. De esta manera podremos evaluar si las dimensiones existentes se pueden reutilizar, también con dimensiones de formulario que pertenecen a otros modelos de negocio, lo que será excelente para optimizar el trabajo de diseño.

Desde la perspectiva del modelo, normalmente los vemos asociados con múltiples tablas de hechos, lo que permite a las organizaciones optimizar el tiempo para realizar cambios y agregar más campos contextuales para usar el modelo. Construya modelos y reutilice dimensiones entre bastidores, ya que esta nueva información puede ser necesaria en cualquier momento para satisfacer el análisis presentado.

Las dimensiones también cambian con el tiempo, por lo que los propietarios de perfiles y los respectivos equipos deben crear un conjunto de políticas para manejar los cambios de dimensiones definidos para cada uno. El propósito de crear una estrategia es delinear el impacto de los cambios en estas áreas dimensionales en el modelo dimensional construido. Los tipos de dimensiones SCD disponibles son:

1. SCD tipo 0, para valores estáticos
2. SCD tipo 1, para valores sobrescritos
3. SCD tipo 2, para la adición de una nueva fila
4. SCD tipo 3, para la adición de un nuevo atributo

Fact tables.

Una fact table contiene medidas numéricas producidas por un evento de medición operacional en el mundo real. A nivel más simple, una fila de una fact table corresponde a una medición de un evento. En adición a las medidas que también son llamadas métricas, una fact table también contiene llaves foráneas que corresponden a las dimensiones asociadas, así como llaves degeneradas. Entre los tipos de Fact table que existen podemos mencionar:

a) Fact Table Transaccional

Una fila en la tabla de hechos de transacciones corresponde a una medición de un evento en un momento dado en el tiempo y el espacio. La precisión de las tablas de hechos de transacciones debe ser lo más baja posible, lo que permitirá cálculos más precisos y un mejor uso de los datos. Solo cuando se midan los eventos habrá entradas en la tabla de hechos de la transacción, y esas entradas deben coincidir con el nivel de granularidad seleccionado.

b) Fact Table de Snapshot periódicos

Una fila en una tabla de hechos instantánea periódica resume muchas mediciones de eventos que ocurrieron durante un período de tiempo estándar, como un día, una semana o un mes. La granularidad son períodos, no transacciones individuales. Las tablas de hechos periódicas instantáneas suelen contener muchas mediciones porque se permiten todas las mediciones de eventos que cumplan con el nivel de precisión.

c) Fact Table de Snapshot acumulativos

Una fila en la tabla de hechos instantáneos acumulativos recopila mediciones de eventos que ocurrieron durante el tiempo estimado entre el inicio y el final del proceso. Puede utilizar las siguientes tablas de hechos para modelar procesos con puntos iniciales, intermedios y finales definidos, como el procesamiento de pedidos o el proceso de reclamaciones.

Modelo de Estrella

Un esquema en estrella consta de múltiples dimensiones relacionadas con una tabla de hechos, por lo que un esquema en estrella representa un proceso de negocio. Los modelos estrella tienen las siguientes características:

- Una Fact table que contiene métricas sobre un proceso de negocio rodeada de dimensiones que proporcionan el contexto temporal de los eventos. • Fácil de entender para los usuarios empresariales
- Simple y simétrico
- Mejorar el rendimiento
- Altamente escalable con nuevas dimensiones y métricas
- veintitrés
- No es para una consulta específica

El un Data Warehouse contendrá tantos modelos en estrella como procesos de negocio se estén analizando, y las dimensiones que componen los modelos en estrella se comparten entre tablas de hechos.

3) CAPITULO II: Análisis y diseño de la propuesta de solución

a) Metodología de Trabajo

Data Warehousing y Business Intelligence

Antes de profundizar en los detalles del modelado dimensional, resulta útil centrarse en los objetivos básicos de la data werehouse y la Business Intelligence. Los objetivos se pueden establecer fácilmente caminando por cualquier organización y escuchando a la dirección de la empresa. Estos temas recurrentes han persistido durante más de tres décadas:

- "Recopilamos muchos datos, pero no tenemos acceso a ellos".
- "Tenemos que cortar y trocear el material de diferentes maneras".
- "Las empresas necesitan un fácil acceso a los datos".
- "Sólo dime qué es importante".
- "Durante toda la reunión estuvimos discutiendo sobre quién tenía los números correctos en lugar de tomar una decisión".
- "Queremos que la gente utilice la información para respaldar decisiones más basadas en hechos".

Empíricamente, estos problemas siguen siendo tan comunes que definen los requisitos básicos para los sistemas DW/BI. Para resolver los problemas anteriores, el sistema DW/BI debe cumplir los siguientes requisitos:

Los sistemas DW/BI deben tener información fácilmente disponible.

El contenido del sistema DW/BI debe ser comprensible. Los datos deben ser intuitivos y obvios para los usuarios empresariales, no sólo para los desarrolladores. La estructura de datos y el marcado deben imitar el proceso de pensamiento y el vocabulario de los usuarios comerciales. Los usuarios empresariales quieren separar y combinar datos analíticos en combinaciones ilimitadas. Las herramientas y aplicaciones de inteligencia empresarial que acceden a los datos deben ser simples y fáciles de usar. ellos también deberían devolver los resultados de la consulta al usuario con tiempos de espera mínimos. Podemos resumir este requisito simplemente diciendo: simple y rápido.

Los sistemas DW/BI deben mostrar información de manera consistente.

Los datos del sistema DW/BI deben ser fiables. Los datos deben recopilarse cuidadosamente de diversas fuentes, limpiarse, garantizarse la calidad y publicarse sólo cuando sea apropiado para las necesidades del usuario. La coherencia también significa utilizar etiquetas y definiciones comunes para el contenido en todas las fuentes del sistema DW/BI. Si dos indicadores de desempeño tienen el mismo nombre, deben tener el mismo significado. Por el contrario, si dos medidas tienen significados diferentes, deberían etiquetarse de manera diferente.

Los sistemas DW/BI deben adaptarse al cambio.

Las necesidades del usuario, las condiciones comerciales, los datos y la tecnología pueden cambiar. Los sistemas DW/BI deben diseñarse para manejar estos cambios inevitables y evitar

errores en los datos o aplicaciones existentes. Cuando la empresa hace nuevas preguntas o agrega datos al repositorio, los datos y aplicaciones existentes no deben modificarse ni interrumpirse. Finalmente, si es necesario cambiar datos descriptivos en el sistema DW/BI, estos cambios deben contabilizarse adecuadamente y hacerse transparentes para los usuarios.

Los sistemas DW/BI deben proporcionar información de manera oportuna.

A medida que los sistemas DW/BI se utilizan cada vez más para la toma de decisiones operativas, es posible que sea necesario transformar los datos sin procesar en información procesable en horas, minutos o incluso segundos. Los equipos de DW/BI y los usuarios empresariales deben tener expectativas realistas sobre el impacto de proporcionar datos cuando no hay tiempo para ejecutarlos, limpiarlos o validarlos.

Los sistemas DW/BI deben ser fortalezas de seguridad que protejan los activos de información.

La joya de la corona de la información de una organización se almacena en un almacén de datos. Como mínimo, el almacén probablemente tenga información sobre lo que desea comprar, a quién y a qué precio, y esa información, en las manos equivocadas, puede arruinar las piezas. Los sistemas DW/BI deben controlar eficazmente el acceso a la información confidencial de la organización.

Los sistemas DW/BI deberían servir como base autorizada y fiable para una mejor toma de decisiones.

El almacén de datos debe tener los datos correctos para la copia de seguridad para la toma de decisiones. Los resultados clave de un sistema DW/BI son decisiones tomadas con base en la evidencia analítica proporcionada; estas decisiones proporcionan el impacto comercial y el valor del sistema DW/BI. La etiqueta original anterior a DW/BI sigue siendo la mejor descripción de lo que está diseñando: un sistema de apoyo a las decisiones.

Una empresa debe adoptar un sistema DW/BI para considerarse exitosa.

No importa si crea una solución elegante con los mejores productos y plataformas. Si la empresa no adopta el entorno DW/BI y lo utiliza activamente, la implementación fracasará. A diferencia de los sistemas transaccionales donde los usuarios empresariales no tienen más opción que utilizar el nuevo sistema, el uso de DW/BI es a veces opcional. Si un sistema DW/BI es una fuente "fácil y rápida" de información procesable, los usuarios empresariales lo utilizarán.

Si bien todos los requisitos de esta lista son importantes, los dos últimos son los más críticos y, lamentablemente, a menudo los que más se pasan por alto. (Kimball, Introducción al modelado dimensional, 2013).

Introducción al Modelado Dimensional

Ahora que comprende los objetivos de un sistema DW/BI, revisemos los conceptos básicos del modelado dimensional. El modelado dimensional se considera ampliamente el método preferido para presentar datos analíticos porque cumple simultáneamente dos requisitos:

- Proporcionar a los usuarios empresariales información comprensible.
- Proporciona un rápido rendimiento de consultas.

El modelado dimensional es una técnica de larga data para simplificar bases de datos. Las organizaciones de TI, los consultores y los usuarios de negocios han gravitado naturalmente durante mucho tiempo hacia estructuras dimensionales simples que satisfacen la necesidad humana básica de simplicidad. La simplicidad es esencial porque garantiza que los usuarios puedan comprender fácilmente los datos y permite que el software se mueva y entregue resultados de forma rápida y eficiente.

La capacidad de visualizar algo tan abstracto como un conjunto de datos de una manera concreta y tangible es el secreto de la inteligibilidad. Un modelo de datos simple al principio puede seguir siendo simple al final del diseño. Es casi seguro que un modelo que comienza siendo complejo terminará siendo demasiado complejo, lo que dará como resultado un rendimiento lento de las consultas y desanimará a los usuarios empresariales.

Aunque los modelos dimensionales a menudo se crean en sistemas de gestión de bases de datos relacionales, son fundamentalmente diferentes de los modelos de tercera forma normal (3NF), cuyo objetivo es evitar la duplicación de datos. La estructura 3NF normalizada divide los datos en varias entidades separadas, cada una de las cuales se convierte en una tabla relacional. Una base de datos de pedidos de ventas puede comenzar con un solo registro por línea de pedido, pero convertirse en un diagrama de araña tan complejo como un modelo 3NF, posiblemente compuesto por cientos de tablas estandarizadas.

La industria a veces se refiere al modelo 3NF como modelo entidad-relación (ER). Un diagrama de entidad-relación (diagrama ER o ERD) es un dibujo que representa las relaciones entre tablas. Tanto el modelo 3NF como el dimensional pueden representarse mediante ERD porque constan de tablas relacionales vinculadas; la diferencia entre 3NF y los modelos dimensionales es el grado de normalización. Dado que ambos tipos de modelos pueden representarse como ERD, evitamos referirnos a los modelos 3NF como modelos ER; en cambio, nos referimos a ellos como modelos estandarizados para reducir la confusión.

Las estructuras 3NF normalizadas son útiles en el procesamiento operativo porque las transacciones de actualización o inserción tocan el almacenamiento en un solo lugar. Sin embargo, el modelo estandarizado es demasiado complejo para consultas de BI y los usuarios no pueden entender, navegar o recordar un modelo estandarizado que parece un mapa del sistema de autopistas de Los Ángeles. Además, la mayoría de los sistemas de gestión de bases de datos relacionales no pueden consultar de manera eficiente modelos normalizados; la complejidad de las consultas impredecibles de los usuarios abruma al optimizador de la base de datos, lo que da como resultado un rendimiento de las consultas catastrófico. El uso de modelos normalizados en el dominio representacional de DW/BI socava la recuperación de datos intuitiva y eficiente. Afortunadamente, el modelado dimensional resuelve el problema de los esquemas demasiado complejos en el área de demostración.

Un modelo dimensional contiene la misma información que un modelo normalizado, pero empaqueta los datos en un formato que brinda comprensión al usuario, rendimiento de consultas y resiliencia al cambio. (Kimball, Dimensional Modeling Introduction, 2013).

Esquema de Estrella

En las bases de datos usadas para data warehousing, un esquema en estrella es un modelo de datos que tiene una tabla de hechos (Fact Table) que contiene los datos para el análisis, rodeada

de las tablas de dimensiones, como se muestra en la Figura 1. Este aspecto, de tabla de hechos más grande rodeada de radios o tablas más pequeñas es lo que asemeja a una estrella, dándole nombre a este tipo de construcciones.

Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que, en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales.

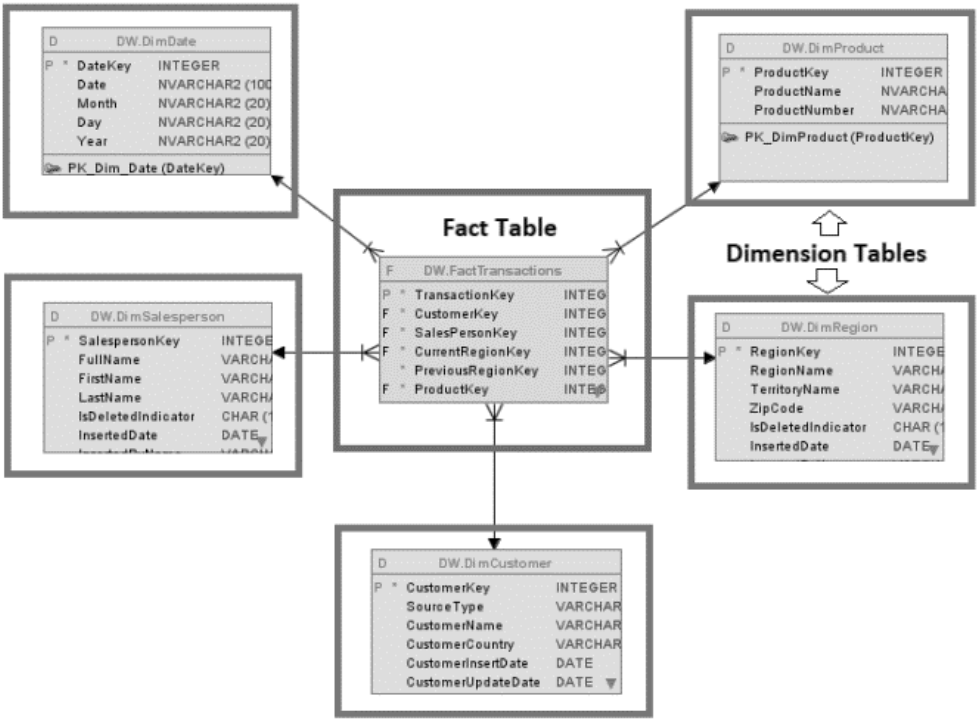


Figura 8. Esquema de Estrella (Model General)

Los dos componentes principales de un plan estrella se describen a continuación.

Tabla de Hechos (Fact Table)

Las tablas de hechos en el modelo dimensional almacenan métricas de rendimiento derivadas de eventos en los procesos comerciales de la organización. Debería intentar almacenar las métricas de bajo nivel generadas por los procesos de negocio en un modelo unidimensional, como se muestra en la Figura 2. Como los datos de medición son, con diferencia, el conjunto de datos más grande, no deben copiarse en múltiples ubicaciones en múltiples funciones de la organización empresarial.

Cada fila de la tabla de hechos corresponde a un evento de medición. Los datos de cada fila tienen un cierto nivel de detalle (llamado granularidad), por ejemplo, una fila para cada producto vendido en una transacción de venta. Uno de los principios fundamentales del modelado dimensional es que todas las filas de dimensiones en una tabla de hechos deben tener la misma precisión. La construcción de tablas de hechos estrictamente con el mismo nivel de detalle garantiza que las mediciones no se enumeren incorrectamente.

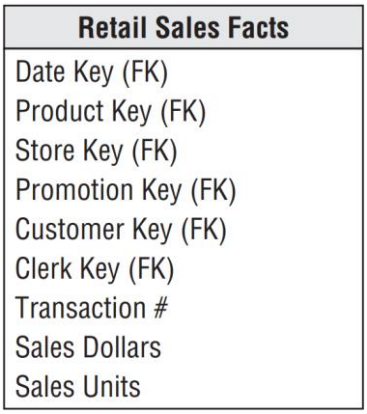


Figura 9. Tabla de Hechos (Kimball, Fact Tables for Measurements, 2013)

La idea de que un evento de medición en el mundo físico tiene una relación uno a uno con una sola fila en la tabla de hechos correspondiente es un principio fundamental del modelado dimensional. Todo lo demás se construye sobre esta base.

Tablas de dimensiones (DimTable)

Las tablas de dimensiones son un complemento completo de las tablas de hechos. Las tablas de dimensiones contienen contextos textuales relacionados con eventos de medición de procesos de negocio. Describen el "qué, qué, dónde, cuándo, cómo y por qué" asociado con un evento.

Las tablas de dimensiones suelen tener muchas columnas o atributos. No es raro que las tablas de dimensiones tengan entre 50 y 100 atributos, como se muestra en la Figura 3; Por supuesto, algunas tablas de dimensiones tienen sólo unos pocos atributos. Las tablas de dimensiones suelen tener menos filas que las tablas de hechos, pero pueden ser anchas y contener muchas columnas de texto grandes. Cada dimensión está definida por una única clave primaria que sirve como base para la integridad referencial de cualquier tabla de hechos con la que esté asociada.

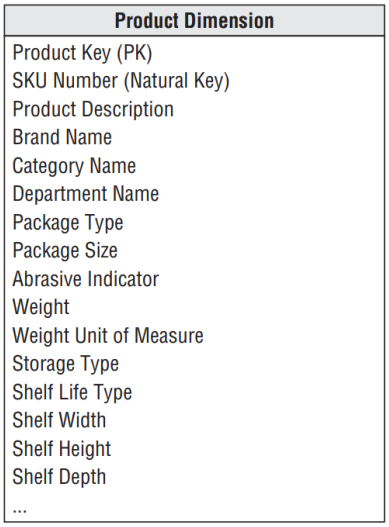


Figura 10. Tablas de Dimensiones (Kimball, Dimension Tables for Descriptive Context, 2013)

Los atributos de la tabla de dimensiones juegan un papel crucial en los sistemas DW/BI. Debido a que los atributos de dimensión son la fuente de casi todas las restricciones y etiquetas de los informes, los atributos de dimensión son fundamentales para que un sistema DW/BI sea utilizable y comprensible. Las propiedades deben ser palabras reales, no abreviaturas ocultas. Debería intentar reducir el uso de código en las tablas de dimensiones reemplazando el código con atributos de texto más detallados.

Tabla de Hechos y Dimensiones unidos en un esquema de estrella

Ahora que comprende las tablas de hechos y las tablas de dimensiones, es hora de juntar los componentes básicos para crear un modelo dimensional. Cada proceso de negocio está representado por un modelo dimensional que consta de una tabla de hechos que contiene las medidas numéricas de un evento rodeada por un conjunto de tablas de dimensiones que contienen el contexto real cuando ocurrió el evento; como se muestra en la Figura 4.

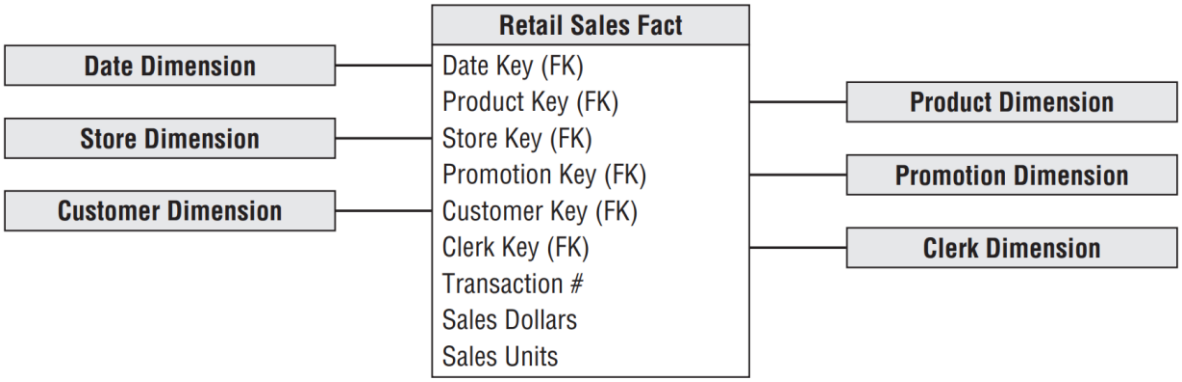


Figura 11. Tablas de hechos y dimensiones en un modelo dimensional. (Kimball, Facts and Dimensions Joined in a Star Schema, 2013)

Lo primero que hay que notar sobre el esquema dimensional es su simplicidad y simetría. Claramente, los usuarios empresariales se benefician de la simplicidad, ya que los datos son más fáciles de entender y navegar. Además, reducir la cantidad de tablas y utilizar descripciones significativas de la empresa facilita la navegación y reduce la posibilidad de errores.

La simplicidad del modelo dimensional también tiene beneficios de rendimiento. El optimizador de la base de datos puede manejar estos modelos simples de manera más eficiente con menos conexiones. El motor de base de datos puede hacer suposiciones sólidas acerca de restringir primero las tablas de dimensiones indexadas en altura y luego atacar la tabla de hechos de una sola vez utilizando el producto cartesiano de las claves de la tabla de dimensiones que satisfacen las restricciones del usuario. Sorprendentemente, con este enfoque, el optimizador puede evaluar cualquier unión de n vías a la tabla de hechos en un solo recorrido utilizando el índice de la tabla de hechos.

Arquitectura de Kimball: Data Warehousing y Business Intelligence

Debe comprender la importancia estratégica de cada componente para evitar confusión sobre su papel y función. Un entorno DW/BI debe considerar cuatro elementos separados y distintos:

sistemas fuente operativos, sistemas ETL, áreas de presentación de datos y aplicaciones de inteligencia empresarial, como se muestra en la Figura 5.

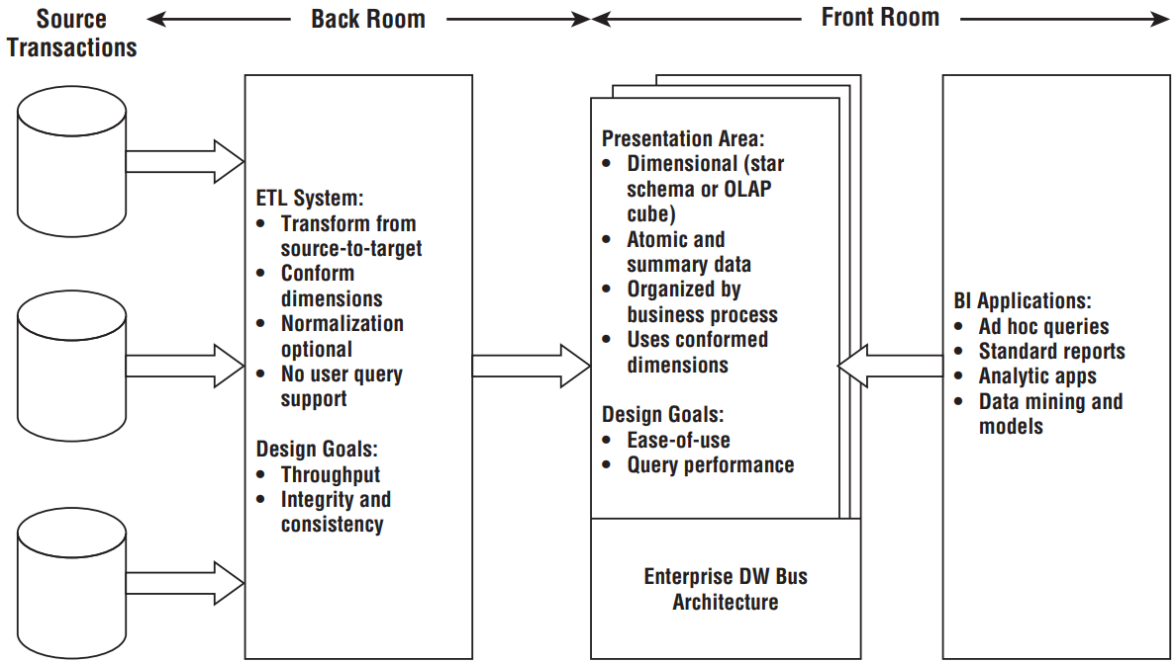


Figura 12. Elementos centrales de la arquitectura Kimball DW/BI. (Kimball, Kimball's DW/BI Architecture, 2013)

Sistema de extracción, transformación y carga

Un sistema de extracción, transformación y carga (ETL) en un entorno DW/BI consta de un espacio de trabajo, estructuras de datos instanciadas y un conjunto de procedimientos. El sistema ETL es todo lo que se encuentra entre los sistemas fuente operativos y el área de presentación DW/BI. Este elemento es esencial para el rompecabezas general de un sistema DW/BI.

La adquisición es el primer paso en el proceso de llevar datos a un entorno de almacén de datos. Extracción significa leer y comprender los datos de origen y copiar los datos requeridos en el sistema ETL para operaciones posteriores. En este punto, los datos pertenecen al almacenamiento de datos.

Una vez que los datos se extraen en el sistema ETL, son posibles muchas transformaciones, como limpiar los datos (corregir errores tipográficos, resolver conflictos de dominio, manejar elementos faltantes o analizarlos en un formato estándar), fusionar datos de múltiples fuentes y eliminarlos. Los datos están respaldados. Los sistemas ETL agregan valor a los datos al realizar estas tareas de limpieza y comparación, transformando y mejorando los datos. Además, estas actividades se pueden diseñar para crear metadatos de diagnóstico que, en última instancia, conduzcan al rediseño de los procesos de negocio para mejorar la calidad de los datos en los sistemas de origen con el tiempo.

El último paso en el proceso ETL es estructurar físicamente y cargar los datos en el modelo dimensional objetivo del área de visualización. Estos subsistemas son esenciales porque la tarea principal de un sistema ETL es transferir datos a tablas de dimensiones y hechos. Muchos de los subsistemas definidos se centran en el manejo de la tabla de dimensiones, como asignar una clave sustituta, búsquedas de códigos para proporcionar descripciones apropiadas, dividir o unir columnas para mostrar valores de hechos apropiados o una tercera base de forma normal para conectar dimensiones planas desnormalizadas a la tabla. estructura. Por el contrario, las tablas de hechos suelen ser grandes y su carga requiere mucho tiempo, pero generalmente son fáciles de preparar para el área de presentación. Se notifica al mundo empresarial que se ha publicado nuevo material cuando las dimensiones y tablas de hechos del modelo dimensional se han actualizado, indexado, proporcionado con agregaciones apropiadas y con mayor calidad.

Área de presentación de apoyo a Business Intelligence

El área de presentación de DW/BI es donde los usuarios, los redactores de informes y otras aplicaciones de análisis de BI organizan, almacenan y consultan los datos directamente. Debido a que los sistemas ETL backend son limitados, el área de demostración es el entorno DW/BI en lo que se refiere al negocio; es todo lo que una empresa ve y toca a través de sus herramientas de acceso y aplicaciones de BI.

El área de datos de presentación debe estructurarse utilizando eventos de medición de procesos de negocio. Este enfoque coincide naturalmente con los sistemas operativos de captura de datos de origen. Los modelos dimensionales deben corresponder a los eventos físicos de recolección de datos; no deben estar diseñados para proporcionar informes intradías. Los procesos comerciales de la empresa abarcan los límites funcionales de departamentos y organizaciones.

En otras palabras, debe crear una tabla de hechos para métricas de ventas atómicas en lugar de completar bases de datos de métricas de ventas individuales similares, pero ligeramente diferentes para los equipos de ventas, marketing, logística y finanzas.

Los datos en el área de vista de consultas de un sistema DW/BI deben ser dimensionales, atómicos (completos con agregación de rendimiento avanzada), centrados en los procesos de negocio y ajustarse a la arquitectura del bus del almacén de datos empresarial. Los datos no deben construirse en base a la interpretación de los datos por parte de cada departamento.

b) Descripción de la propuesta de solución

Resultado del Data Profiling

Para el procesamiento de datos, empleamos DATA CLEANER, un programa gratuito que ofrece diversas funcionalidades, entre ellas el procesamiento de archivos CSV. Esta herramienta nos permite analizar la calidad de los datos obtenidos, así como encontrar patrones y supervisar los valores presentes en los mismos. Es importante resaltar que utilizaremos DATA CLEANER para realizar el perfilado de los datos, con el objetivo de verificar la integridad de los dataset investigados por el equipo.

El análisis que se realizó al conjunto de Dataset y en el perfilamiento de los datos de cada uno de los csv se encontraron los siguientes puntos.

Acc_Classified_according_to_Type_of_Weather_Condition_2014_and_2016								
Descripción: La tabla cuenta con 37 registros que son equivalentes a los estados analizados. Estos se desglosan en columnas que conforman la información de clima para los años 2016 y 2014 (Un breve análisis de su estructura demuestra que no se encuentran valores nulos en la composición de este datasets el análisis de la información se desglosa a continuación)								
Column	Row count	Null count	Blank count	Total chart count	Avg Chart	Digit chars	Max chars	Min chars
Fine - Total Acc. - 2014	37	0	0	137	3703	135135	6	1
Fine - Persons Killed - 2014	37	0	0	120	3243	120	5	1
Fine - Persons Injured - 2014	37	0	0	137	3703	135	6	1
Mist/fog - Total Acc. - 2014	37	0	0	93	2514	91	5	1
Mist/fog - Persons Killed - 2014	37	0	0	82	2216	82	4	1
Mist/fog - Persons Injured - 2014	37	0	0	92	2486	90	5	1
Cloudy - Total	37	0	0	94	2541	92	5	1

Acc. - 2014								

Cloudy - Persons Killed - 2014	37	0	0	80	2162	78	4	1
Cloudy - Persons Injured - 2014	37	0	0	92	2486	90	5	1
Light rain - Total Acc. - 2014	37	0	0	104	2811	102	5	1
Light rain - Persons Killed - 2014	37	0	0	86	2324	84	4	1
Light rain - Persons Injured - 2014	37	0	0	104	2811	102	5	1
Heavy rain - Total Acc. - 2014	37	0	0	95	2568	93	5	1
Heavy rain - Persons Killed - 2014 Heavy rain - Persons Injured - 2014	37	0	0	79	2135	79	4	1
Flooding of slipways/rivulers - Total Acc. -	37	0	0	96	2595	94	5	1

2014								

Flooding of slipways/rivulers - Persons Injured - 2014	37	0	0	55	1486	53	3	1
Hail/sleet - Persons Killed - 2014	37	0	0	68	1838	66	4	1
Hail/sleet - Persons Injured - 2014	37	0	0	61	1649	59	4	1
snow - Total Acc. - 2014	37	0	0	52	1405	50	3	1
snow - Persons Killed - 2014	37	0	0	60	1622	58	4	1
snow - Persons Injured - 2014	37	0	0	62	1676	60	4	1
Strong wind - Total Acc. - 2014	37	0	0	56	1514	54	3	1

Strong wind - Persons Killed - 2014	37	0	0	63	1703	61	4	1
Strong wind - Persons Injured - 2014	37	0	0	75	2027	73	4	1
Dust storm - Total	37	0	0	63	1703	63	4	1

Acc. - 2014								
Dust storm - Persons Killed - 2014	37	0	0	72	1946	70	4	1
Dust storm - Persons Injured - 2014	37	0	0	75	2027	73	4	1
Very hot - Total Acc. - 2014	37	0	0	64	1730	62	4	1
Very hot - Persons Killed - 2014	37	0	0	75	2027	73	4	1
Very hot - Persons Injured - 2014	37	0	0	89	2405	87	5	1
Very cold - Total Acc. - 2014	37	0	0	79	2135	77	4	1

Very cold - Persons Killed - 2014	37	0	0	90	2432	88	5	1
Very cold - Persons Injured - 2014	37	0	0	86	2324	84	5	1
Other extraordinary weather condition - Total	37	0	0	76	2054	74	5	1

Acc. – 2014								
Other extraordinary weather condition - Persons Killed - 2014	37	0	0	88	2378	86	5	1
Other extraordinary weather condition - Persons Injured - 2014	37	0	0	103	2784	101	5	1
Fine/Clear - Total Accidents - 2016	37	0	0	141	3811	141	6	1

Fine/Clear - Persons Killed - 2016	37	0	0	123	3324	123	6	1
Fine/Clear - Persons Injured - 2016	37	0	0	140	3784	140	6	1
Mist/ Foggy - Total Accidents - 2016	37	0	0	95	2568	95	5	1
Mist/ Foggy - Persons Killed - 2016	37	0	0	81	2189	81	4	1
Mist/ Foggy - Persons Injured - 2016	37	0	0	96	2595	96	5	1
Cloudy - Total Accidents - 2016	37	0	0	94	2541	64	5	1
Cloudy - Persons Killed - 2016	37	0	0	78	2108	78	4	1
Cloudy - Persons Injured - 2016	37	0	0	96	2595	96	5	1
Rainy - Total Accidents - 2016	37	0	0	105	2838	105	5	1

Rainy - Persons Killed - 2016	37	0	0	88	2378	88	5	1
Rainy - Persons Injured - 2016	37	0	0	107	2892	107	5	1
Snowfall - Total Accidents - 2016	37	0	0	40	1081	40	2	1
Snowfall - Persons Killed - 2016	37	0	0	38	1027	38	2	1
Snowfall - Persons Injured - 2016	37	0	0	41	1108	41	2	1
Hail/Sleet - Total Accidents - 2016	37	0	0	58	1568	58	4	1
Hail/Sleet - Persons Killed - 2016	37	0	0	51	1378	51	3	1
Hail/Sleet - Persons Injured - 2016	37	0	0	59	1595	59	4	1
Dust Storm - Total Accidents - 2016	37	0	0	68	1838	68	4	1

Dust Storm - Persons Killed - 2016	37	0	0	63	1703	63	4	1
Dust Storm - Persons Injured - 2016	37	0	0	70	1892	70	4	1
Others - Total Acciden ts - 2016	37	0	0	98	2649	98	5	1
Others - Persons Killed - 2016	37	0	0	87	2351	87	5	1
Others - Persons Injured - 2016	37	0	0	97	2622	98	5	1

Acc_clf_acco_to_Road_Cond_2014_and_2016
Descripción: La tabla cuenta con 37 registros que son equivalentes a los estados analizados. Estos se desglosan en columnas que conforman la información del estado de camino para los años 2016 y 2014 (Un breve análisis de su estructura demuestra que no se encuentran valores nulos en la composición de este datasets el análisis de la información se desglosa a continuación)

Column	Row count	Null count	Blank count	Total chart count	Avg Chart	Digit chars	Max chars	Min chars
Surfaced Roads- Accident - 2014	37	0	0	136	1838	134	6	1
Surfaced Roads- Killed - 2014	37	0	0	123	3324	123	5	1
Surfaced Roads- Injured - 2014	37	0	0	140	3784	138	6	1

Metalled Roads-Accident - 2014	37	0	0	113	3054	111	5	1
Metalled Roads- Killed - 2014	37	0	0	101	273	101	5	1
Metalled Roads-Injured - 2014	37	0	0	112	3027	110	5	1
Kutcha Roads-Accident - 2014	37	0	0	104	2811	102	5	1
Kutcha Roads- Killed - 2014	37	0	0	89	2405	89	5	1
Kutcha Roads-Injured - 2014	37	0	0	102	2757	100	5	1
Dry roadAccident - 2014	37	0	0	143	3865	141	6	1
Dry road-Killed - 2014	37	0	0	128	3459	128	6	1
Dry roadInjured - 2014	37	0	0	144	3892	142	6	1
Wet roadAccident - 2014	37	0	0	110	2973	108	5	1
Wet road-Killed - 2014	37	0	0	101	273	101	5	1
Wet road-Injured - 2014	37	0	0	113	3054	111	5	1
Good surfaceAccident - 2014	37	0	0	136	3676	134	6	1

Good surface- Killed - 2014	37	0	0	123	3324	123	5	1
Good surfaceInjured - 2014	37	0	0	138	373	136	6	1
Loose SurfaceAccident - 2014	37	0	0	102	3757	100	5	1
Loose Surface- Killed - 2014	37	0	0	87	2351	87	4	1
Loose SurfaceInjured - 2014	37	0	0	102	2757	100	5	1
Rutted/Pot holesAccident - 2014	37	0	0	79	2135	77	5	1
Rutted/Pot holes- Killed - 2014	37	0	0	68	1838	68	4	1
Rutted/Pot holes-Injured - 2014	37	0	0	76	2054	74	5	1
Road under repair/constr uctionAccident - 2014	37	0	0	83	2243	81	5	1
Road under repair/constr uction- Killed - 2014	37	0	0	74	2	74	4	1
Road under repair/constr uctionInjured - 2014	37	0	0	81	2189	79	5	1

Corrugated/ Wavy roadAccident - 2014	37	0	0	73	1973	71	4	1
Corrugated/ Wavy road- Killed - 2014	37	0	0	64	173	64	4	1
Corrugated/ Wavy road- Injured - 2014	37	0	0	74	2	72	4	1
Slippery surfaceAccident - 2014	37	0	0	74	2108	76	4	1
Slippery surface- Killed - 2014	37	0	0	66	1784	66	4	1
Slippery surfaceInjured - 2014	37	0	0	78	2108	76	4	1
Snowy- Accident - 2014	37	0	0	61	1649	59	4	1
Snowy- Killed - 2014	37	0	0	48	1297	48	3	1
Snowy- Injured - 2014	37	0	0	60	1622	58	4	1
Muddy- Accident - 2014	37	0	0	71	1919	69	4	1
Muddy- Killed - 2014	37	0	0	61	1649	61	4	1
Muddy- Injured - 2014	37	0	0	75	2027	73	4	1
Oily-Accident - 2014	37	0	0	62	1676	60	4	1
Oily- Killed - 2014	37	0	0	50	1351	50	5	1

Oily-Injured - 2014	37	0	0	61	1649	59	5	1
Speed breakerAccident - 2014	37	0	0	85	2297	83	5	1
Speed breaker-Killed - 2014	37	0	0	66	1784	66	5	1
Steep InclineInjured - 2014	37	0	0	83	2243	81	5	1
Hump-Accidents - 2014	37	0	0	104	2811	102	4	1
Hump-Killed - 2014	37	0	0	103	2784	103	4	1
Hump-Injured - 2014	37	0	0	84	227	84	5	1
Dip-Accidents - 2014	37	0	0	104	2811	104	5	1
Dip-Killed - 2014	37	0	0	142	3838	140	6	1
Dip-Injured - 2014	37	0	0	126	3405	126	6	1
Pucca road (Normal Road) - Number of Accidents - 2016	37	0	0	137	3757	137	6	1
Pucca road (Normal Road) - Persons Killed - 2016	37	0	0	102	2757	100	5	1
Pucca road (Normal Road) - Persons Injured - 2016	37	0	0	88	2378	88	5	1

Kutcha road (Normal Road) -	37	0	0	104	2811	102	5	1

Number of Accidents - 2016								
Kutcha road (Normal Road) - Persons Killed - 2016	37	0	0	89	2405	89	5	1
Kutcha road (Normal Road) - Persons Injured - 2016	37	0	0	76	2054	76	4	1
Pot Holes - Number of Accidents - 2016	37	0	0	90	2432	90	5	1
Pot Holes - Persons Killed - 2016	37	0	0	82	2216	82	5	1
Pot Holes - Persons Injured - 2016	37	0	0	71	1919	71	4	1
Speed Breakers - Number of Accidents - 2016	37	0	0	86	2324	86	5	1
Speed Breakers - Persons Killed - 2016	37	0	0	80	2162	80	5	1
Speed Breakers - Persons Injured - 2016	37	0	0	73	1973	73	4	1

Sharp Curve - Number of Accidents - 2016	37	0	0	83	2243	83	5	1
Sharp Curve - Persons Killed - 2016	37	0	0	140	3784	140	6	1
Sharp Curve - Persons Injured - 2016	37	0	0	123	3324	123	6	1
Steep Gradient - Number of Accidents - 2016	37	0	0	140	3784	140	6	1
Steep Gradient - Persons Killed - 2016	37	0	0	103	2784	103	5	1
Steep Gradient - Persons Injured - 2016	37	0	0	83	2216	82	4	1

only_road_accidents_data_month2								
Descripción: La tabla cuenta con 490 registros que son equivalentes a los meses analizados. Estos se desglosan en columnas que conforman la información en meses de los años 2001 al 2014 por todos los estados (Un breve análisis de su estructura demuestra que no se encuentran valores nulos en la composición de este Dataset el análisis de la información se desglosa a continuación)								
Column	Row count	Null count	Blank count	Total chart count	Avg Chart	Digit chars	Max chars	Min chars
JANUARY	490	0	0	1365	2786	1365	4	1
FEBRUARY	490	0	0	1359	2773	1359	4	1
MARCH	490	0	0	1370	2796	1370	4	1
APRIL	490	0	0	1366	2788	1366	4	1

MAY	490	0	0	1367	279	1367	4	1
JUNE	490	0	0	1364	2784	1364	4	1
JULY	490	0	0	1358	2771	1358	4	1
AUGUST	490	0	0	1348	2751	1348	4	1
SEPTEMBER	490	0	0	1349	2753	1349	4	1
OCTOBER	490	0	0	1361	2778	1361	4	1
NOVEMBER	490	0	0	1360	2776	1360	4	1
DECEMBER	490	0	0	1375	2806	1365	4	1
TOTAL	490	0	0	1865	3806	1865	5	1

only_road_accidents_data3								
Descripción: La tabla cuenta con 490 registros que son equivalentes a las horas por las cuales suceden los accidentes. Estos se desglosan en columnas que conforman la información en meses de los años 2001 al 2014 por todos los estados esto es un total de todos (Un breve análisis de su estructura demuestra que no se encuentran valores nulos en la composición de este datasets el análisis de la información se desglosa a continuación)								
Column	Row count	Null count	Blank count	Total chart count	Avg Chart	Digit chars	Max chars	Min chars
0-3	490	0	0	1275	2602	1275	4	1
3-6	490	0	0	1318	269	1318	4	1
6-9	490	0	0	1430	2918	1430	4	1
9-12	490	0	0	1493	3047	1493	5	1
12-15	490	0	0	1483	3027	1483	5	1
15-18	490	0	0	1499	3059	1499	5	1
18-21	490	0	0	1483	3027	1483	5	1
21-24	490	0	0	1404	2865	1404	5	1
total	490	0	0	1865	3806	1865	5	1

Road_Accidents_2017-Annuxure_Tables_3								
Descripción: La tabla cuenta con 37 registros que son equivalentes sumatorias de los accidentes y personas muertas. Estos se desglosan en columnas que conforman la información en años del 2014 al 2017 por todos los estados (Un breve análisis de su estructura demuestra que no se encuentran valores nulos en la composición de este datasets el análisis de la información se desglosa a continuación)								
Column	Row count	Null count	Blank count	Total chart count	Avg Chart	Digit chars	Max chars	Min chars
State/UTwise Total Number of Persons Killed in Road Accidents during - 2014	37	0	0	132	3568	132	6	1
State/UTwise Total Number of Persons Killed in Road Accidents during - 2015	37	0	0	131	3541	131	6	1
State/UTwise Total Number of Persons Killed in Road Accidents during - 2016	37	0	0	130	3514	130	6	1
State/UTwise Total Number of Persons Killed in Road Accidents during - 2017	37	0	0	132	3568	132	6	1

Share of States/UTs in Total Number of Persons Killed in Road Accidents - 2014	37	0	0	103	2784	72	4	1

Share of States/UTs in Total Number of Persons Killed in Road Accidents - 2015	37	0	0	95	2568	68	4	1
Share of States/UTs in Total Number of Persons Killed in Road Accidents - 2016	37	0	0	97	2622	69	4	1
Share of States/UTs in Total Number of Persons Killed in Road Accidents - 2017	37	0	0	99	2676	70	4	1
Total Number of Persons Killed in Road Accidents Per Lakh Population -	37	0	0	114	3081	81	4	1

2014								
Total Number of Persons Killed in Road Accidents Per Lakh Population - 2015	37	0	0	120	3243	86	4	1
Total Number of Persons Killed in Road Accidents Per Lakh Population - 2016	37	0	0	117	3162	85	4	1
Total Number of Persons Killed in Road Accidents Per Lakh Population - 2017	37	0	0	125	3378	88	4	1
Total Number of Persons Killed in Road Accidents per 10,000	37	0	0	112	3027	78	4	1

Vehicles - 2014								
Total Number of Persons Killed in Road Accidents per 10,000 Vehicles - 2015	37	0	0	108	2919	75	4	1
Total Number of Persons Killed in Road Accidents per 10,000 Vehicles - 2016	37	0	0	111	3	76	4	1
Total Number of Persons Killed in Road Accidents per 10,000 Km of Roads - 2014	37	0	0	164	4432	134	6	1
Total Number of Persons Killed in Road Accidents	37	0	0	170	4595	136	6	1

per 10,000 Km of Roads - 2015								
Total Number of	37	0	0	176	4757	140	6	1
Persons Killed in Road Accidents per 10,000 Km of Roads - 2016								

Road_Accidents_2017-Annuxure_Tables_4								
Descripción: La tabla cuenta con 37 registros que son equivalentes sumatorias de los accidentes personas muertas. Estos se desglosan en columnas que conforman la información en años del 2014 al 2017 por todos los estados (Un breve análisis de su estructura demuestra que no se encuentran valores nulos en la composición de este datasets el análisis de la información se desglosa a continuación)								
Column	Row count	Null count	Blank count	Total chart count	Avg Chart	Digit chars	Max chars	Min chars
State/UTwise Total Number of Persons Injured in Road Accidents during - 2014	37	0	0	149	4027	149	6	1
State/UTwise Total Number of Persons Injured in Road Accidents during - 2015	37	0	0	149	4027	149	6	1

State/UTwise Total Number of Persons Injured in Road Accidents during - 2016	37	0	0	148	4	148	6	1
State/UTwise Total Number of Persons Injured in Road Accidents during - 2017	37	0	0	104	2757	73	4	1

Share of States/UTs in Total Number of Persons Injured in Road Accidents - 2014	37	0	0	102	2703	72	4	1
Share of States/UTs in Total Number of Persons Injured in Road Accidents - 2015	37	0	0	100	2865	71	4	1

Share of States/UTs in Total Number of Persons Injured in Road Accidents - 2016	37	0	0	106	2811	74	4	1
Share of States/UTs in Total Number of Persons Injured in Road Accidents - 2017	37	0	0	143	3865	106	5	1
Total Number of Persons Injured in Road Accidents Per Lakh Population - 2014	37	0	0	141	3811	105	5	1
Total Number of Persons	37	0	0	134	3622	102	5	1

Injured in Road Accidents Per Lakh Population - 2015								

Total Number of Persons Injured in Road Accidents Per Lakh Population - 2016	37	0	0	142	3838	106	5	1
Total Number of Persons Injured in Road Accidents Per Lakh Population - 2017	37	0	0	135	3649	101	4	1
Total Number of Persons injured in Road Accidents per 10,000 Vehicles - 2014	37	0	0	133	3595	99	4	1
Total Number of Persons injured in Road Accidents per 10,000 Vehicles - 2015	37	0	0	132	3568	98	4	1

Total Number of Persons injured in Road Accidents per	37	0	0	197	5327	163	6	1
10,000 Vehicles - 2016								
Total Number of Persons injured in Road Accidents per 10,000 Km of Roads - 2014	37	0	0	199	5378	164	6	1
Total Number of Persons injured in Road Accidents per 10,000 Km of Roads - 2015	37	0	0	190	5135	158	6	1
Total Number of Persons injured in Road Accidents per 10,000 Km of Roads - 2016	37	0	0	152	5623	125	5	1

Tabla 2. Resultados del Data Profiling

En base a la información recabada de los CSV y con el data Profiling se propuso el siguiente diagrama dimensional.

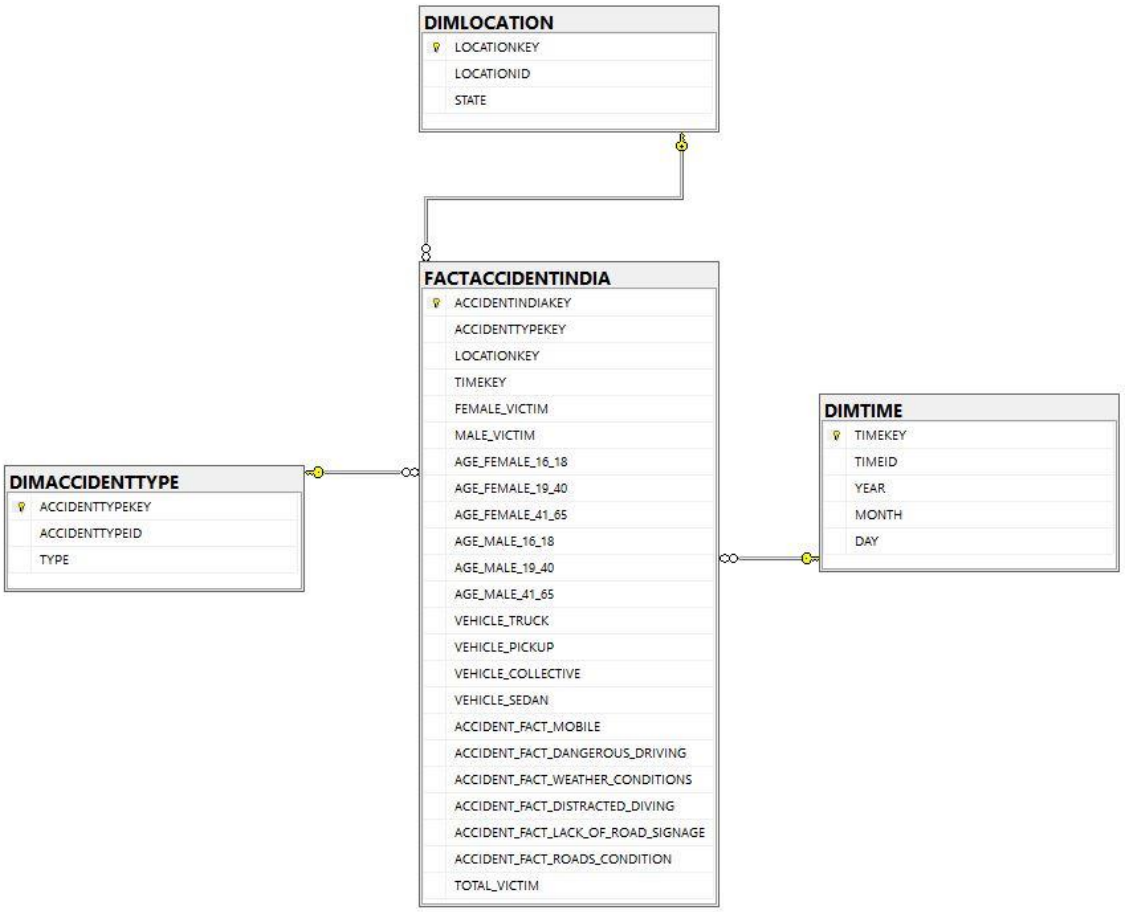


Figura 13. Diagrama de Estrella Propuesto

Especificación de necesidades analíticas que el modelo propuesto solventará

Nuestro modelo dimensional está basado en el data set de los accidentes de tránsito en la india, las necesidades que este solventara son las siguientes:

- El total de accidentes de tránsito durante un periodo definido.
- La cantidad de lesionados por género y edad en los accidentes de tránsito.
- La cantidad de Fallecidos por género y edad en los accidentes de tránsito.
- La frecuencia de accidentes por ubicación definida.
- El total de accidentes por tipo de vehículo (camioneta, sedan, microbús)

Estas son solo algunas de las preguntas que podríamos responder, se espera que el modelo dimensional permita a los usuarios analíticos poder contestar más tipos de preguntas.

c) Descripción de la tecnología a usar

Talen Open Studio

Talend Open Studio (TOS) es una suite que aporta un conjunto muy complejo, variado y completo de herramientas para llevar a cabo la integración de datos que se ofrece en una versión de código libre (open source). Precisamente por ello, esta es una de las herramientas de integración ETL (extract, transform, load) más utilizadas dentro del mundo Big Data; es más, es la cuarta en la lista después de Informática Powercenter, IBM InfoSphere Datastage y Oracle Data Integrator (ODI).

Por otra parte, esta suite cuenta con un Comunity Edition (CE) totalmente funcional. Además, podrás utilizar una gran cantidad de componentes (más o menos 900) para llevar a cabo una gestión de datos personalizada.

Talend ofrece una colección de componentes genéricos destinados a los procesos de transformación de datos entre los que se incluyen las funcionalidades de normalizar, denormalizar, extraer o insertar campos en varios formatos, separar filas, agregar y ordenar filas, convertir tipos, etc.

Asimismo, ofrece conjuntos de componentes de procesamiento de datos que son específicos para diferentes tecnologías. Estos componentes tienen la finalidad de ejecutar dichas transformaciones de la forma más eficiente posible.

Amazon S3

Antes de que profundices en las características de Amazon S3, es relevante que conozcas que el sistema de Amazon Simple Storage Service se define como una herramienta de AWS que se encarga del almacenamiento de objetos, ofreciendo disponibilidad de datos, escalabilidad, alto rendimiento y niveles de seguridad.

Amazon S3 puede implementarse en los procesos de protección y almacenamiento de datos, sin importar su cantidad. Además, este servicio incluye casos de uso como copias de seguridad, análisis de Big Data, aplicaciones móviles, sitios web y demás.
Características de Amazon S3

El sistema de Amazon S3 incluye una serie de características que permiten su funcionamiento, como, por ejemplo, que tiene la capacidad de almacenar y mover datos entre las storage classes de S3.

Otras de las características de Amazon S3 es que contribuye en los procesos de configuración y aplicación de controles relacionados con el acceso a datos, protegiéndolos de usuarios sin autorización para ingresar al sistema.

De la misma manera, Amazon S3 puede utilizarse para realizar una revisión del uso de almacenamiento, así como las tendencias referentes a la actividad en la organización del cliente.

Amazon Redshift

La herramienta de Amazon Redshift se define como un servicio que se encarga de la gestión de procesos de operación, escalado y ajuste del almacenamiento de datos.

La opción de Amazon Redshift también se caracteriza por ser un servicio de data warehouse gestionado que cuenta con varios petabytes en la nube.

Características de Amazon Redshift

El servicio de Amazon Redshift incluye una serie de propiedades que caracterizan y permiten su funcionamiento, como, por ejemplo, que facilita la ejecución y el escalado del análisis de todos los datos en poco tiempo, sin que sea necesario la gestión de la infraestructura de almacenamiento.

El diseño de Amazon Redshift se caracteriza, además, por encargarse del control y la administración de las labores del almacenamiento de datos, como es el caso del aprovisionamiento de capacidad, la realización de copias de seguridad para el clúster y su supervisión o la aplicación de parches de actualización, entre otras.

Este servicio también permite realizar la conversión de datos en información en tan solo segundos, eliminando la necesidad de gestión de infraestructura, al tiempo que ofrece seguridad y fiabilidad en sus actividades.

Cabe destacar que el funcionamiento de Amazon Redshift se lleva a cabo gracias al uso del lenguaje de dominio específico SQL, que permite el análisis de datos semiestructurados y estructurados para el almacenamiento de datos, databases operativas y demás, lo que proporciona un alto rendimiento por un bajo costo.

Amazon IAM

AWS Identity and Access Management (IAM) es un servicio web que lo ayuda a controlar de forma segura el acceso a los recursos de AWS. Se utiliza IAM para controlar quién está autenticado (ha iniciado sesión) y autorizado (tiene permisos) para utilizar recursos.
IAM ofrece las siguientes características:

Acceso compartido a la cuenta de AWS

Puede conceder permiso a otras personas para administrar y utilizar los recursos de su cuenta de AWS sin tener que compartir su contraseña o clave de acceso.

Permisos detallados

Puede conceder diferentes permisos a diferentes personas para diferentes recursos. Por ejemplo, puede permitir que algunos usuarios completen el acceso a Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple Storage Service (Amazon S3), Amazon DynamoDB, Amazon Redshift y otros servicios de AWS. En el caso de otros usuarios, puede permitir el

acceso de solo lectura a solo algunos buckets de S3 o conceder permiso para administrar solo algunas instancias EC2 o para tener acceso a la información de facturación, pero nada más.

Acceso seguro a los recursos de AWS para aplicaciones que se ejecutan en Amazon EC2

Puede utilizar características de IAM para proporcionar de forma segura credenciales para las aplicaciones que se ejecutan en instancias EC2. Estas credenciales proporcionan permisos a la aplicación para obtener acceso a otros recursos de AWS. Entre los ejemplos se incluyen buckets de S3

Multi-Factor authentication (MFA)

Puede agregar una autenticación de dos factores a la cuenta y a los usuarios individuales para mayor seguridad. Con MFA usted o sus usuarios deben proporcionar no solo una contraseña o clave de acceso para trabajar con la cuenta, sino también un código de un dispositivo configurado específicamente. Si ya utiliza una clave de seguridad FIDO con otros servicios y esta tiene una configuración compatible con AWS. Para obtener más información, consulte Configuraciones admitidas para usar las claves de seguridad FIDO.

Power BI

Power BI es un servicio gratuito de análisis de negocio basado en la nube y visualización de datos, de negocio. Esta herramienta de Business Intelligence (BI), incorporada en la suite de productividad Microsoft Office 365, permite controlar la salud de un negocio mediante un dashboard en vivo, crear informes interactivos con Power BI Desktop y acceder a los datos en cualquier lugar con las aplicaciones nativas de móvil.

Actualmente cuenta con más de 5 millones de usuarios y es utilizado por más de 200.000 empresas. Es ampliamente utilizado en agencias de analítica web y empresas especializadas en Business Intelligence.

Power BI es una solución de análisis empresarial basado en la nube, que permite unir diferentes fuentes de datos, analizarlos y presentar un análisis de estos a través de informes y paneles. Con Power BI se tiene de manera fácil acceso a datos dentro y fuera de la organización casi en cualquier dispositivo. Estos análisis pueden ser compartidos por diferentes usuarios de la misma organización; por lo que directivos, financieros, comerciales, etc., pueden disponer de la información del negocio en tiempo real.

Se conforma fundamentalmente de estos componentes:

- Power BI Desktop: aplicación gratuita de escritorio para transformar, visualizar datos y crear informes de los mismos.

- Power BI Service: servicio online (SaaS) con funcionalidad similar a la aplicación desktop y permite publicar informes y configurar la actualización de datos automáticamente para que el personal de la organización tenga los datos actualizados.

- Power BI Mobile: aplicación móvil disponible para Windows, iOS y Android para visualizar informes y que se actualiza automáticamente con los cambios de los datos.

El flujo de trabajo de Power BI sigue el siguiente orden:

1. En Power BI Desktop se obtienen los datos a través de los diferentes conectores existentes.
2. Se cruzan los datos y se generan visualizaciones de datos e informes.
3. Se publican los informes y visualizaciones a través del servicio Power BI.
4. Diferentes usuarios pueden consultar dichos informes a través de apps en sus móviles de Android y iOS.

d) Diagrama arquitectónico de la solución

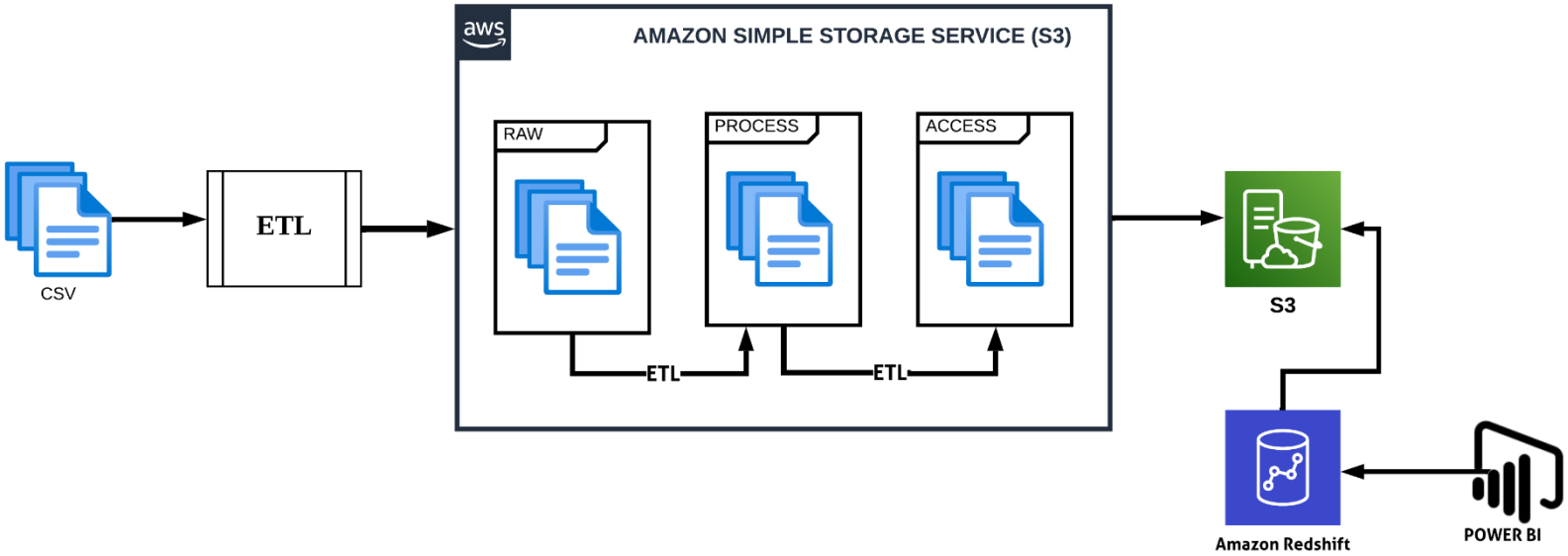


Figura 14. Diagrama Arquitectónico de la Solución

e) Descripción de cada componente de la solución

Procesos ETL

Los procesos ETL es todo lo que se encuentra entre los sistemas fuente operativos y el área de presentación DW/BI. Este elemento es fundamental del rompecabezas general del sistema DW/BI.

- Extraer significa leer y comprender los datos de origen y copiar los datos necesarios en el sistema ETL para su posterior manipulación. En este punto, los datos pertenecen al almacén de datos.

- Las transformaciones, como la limpieza de los datos (corrección de errores ortográficos, resolución de conflictos de dominio, manejo de elementos faltantes o análisis en formatos estándar), combinación de datos de múltiples fuentes y eliminación. Duplicación de datos. El sistema ETL agrega valor a los datos con estas tareas de limpieza y conformidad al cambiar los datos y mejorarlos.

- La carga de datos en los modelos dimensionales de destino del área de presentación. Estos subsistemas son fundamentales, debido a que la misión principal del sistema ETL es entregar los datos a las tablas de dimensiones y hechos.

S3

Utilizaremos S3 para el almacenamiento de los datos procedentes de los archivos CSV. S3 será divido en 3 escenarios, los cuales almacenarán las transformaciones que los datos vayan sufriendo a través de los diferentes procesos ETL que la solución lo requiera. Todo esto a través de una interfaz de servicio web.

S3-RAW DATA

El primer escenario de S3 será RAW DATA, el cuál almacenará la data cruda, directamente de los archivos csv, dichos datos aún no han sufrido ninguna transformación, limpieza o eliminación.

S3-PROCESS DATA

El segundo escenario de S3 será PROCESS DATA, el cuál almacenará los datos que han sufrido algún tipo de transformación, limpieza, eliminación, etc.

S3-ACCESS DATA

El último escenario de S3 será PRESENTATION, el cuál almacenará los datos finales, listos para ser almacenados en el modelo dimensional de la solución.

Amazon RedShift

En Redshift crearemos nuestro modelo dimensional (Tabla de Hecho y Dimensiones), el cuál almacenará los datos provenientes de S3 (Habiendo completado los procesos ELT). Redshift cumplirá la función de base de datos en la nube, la cual cuenta con la capacidad para almacenar y analizar grandes cantidades de registros o fuentes de datos.

Power BI

Power BI será el pasó final del proyecto. Cumple la función de Business Intelligence. Estará conectado con Redshift. Será el encargado de presentar todos los datos de manera gráfica. El objetivo principal es usar los datos para mejorar la toma de decisiones.

4) CAPITULO III: Estrategia de implementación de propuesta de solución

a) Estrategia de implementación

Hemos definido que el producto a entregar será una implementación que se integrará a los sistemas gestores de tránsito de transportes de la India. Esta solución al integrarse respaldará la toma de decisiones. Los resultados más importantes de un sistema DW/BI son las decisiones que se toman con base en la evidencia analítica presentada; estas decisiones brindan un impacto y el valor atribuible al sistema DW/BI. La etiqueta original anterior a DW/BI sigue siendo la mejor descripción de lo que está diseñando: un sistema de soporte de decisiones. Una vez que hemos presentado el diseño de la solución, consideramos que los pasos a seguir para su correspondiente implementación son:

CSV

Verificar la ubicación del archivo de datos Origen los cuales componen el CSV con la información de los accidentes de tránsito en la india.

```
STATE_UT,YEAR,MONTH,DAY,ACCIDENT_TYPE,FEMALE_VICTIM,MALE_VICTIM,AGE_FEMALE_16_18,AGE_
A & N Islands,2007,JANUARY,MONDAY,SERIOUS,2,3,1,0,1,2,0,1,1,0,1,3,4,1,0,0,0,0,5
A & N Islands,2007,JANUARY,TUESDAY,SERIOUS,1,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,1
A & N Islands,2007,JANUARY,WEDNESDAY,MILD,1,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1
A & N Islands,2007,JANUARY,THURSDAY,MILD,0,1,0,0,0,0,1,0,0,0,1,0,1,0,0,0,0,0,1
A & N Islands,2007,JANUARY,FRIDAY,MODERATE,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,JANUARY,SATURDAY,SERIOUS,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,JANUARY,SUNDAY,MODERATE,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,FEBRUARY,MONDAY,MODERATE,5,8,0,1,4,6,0,2,7,4,1,1,10,0,3,0,0,0,13
A & N Islands,2007,FEBRUARY,TUESDAY,SERIOUS,1,6,0,1,0,4,2,0,0,2,2,3,2,0,0,0,0,5,7
A & N Islands,2007,FEBRUARY,WEDNESDAY,MODERATE,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,FEBRUARY,THURSDAY,MILD,1,0,1,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1
A & N Islands,2007,FEBRUARY,FRIDAY,MILD,2,0,1,0,1,0,0,0,0,0,0,2,0,1,0,1,0,0,2
A & N Islands,2007,FEBRUARY,SATURDAY,MILD,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,FEBRUARY,SUNDAY,MODERATE,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,MARCH,MONDAY,MODERATE,4,8,4,0,0,4,1,3,7,4,0,1,5,1,0,4,2,0,12
A & N Islands,2007,MARCH,TUESDAY,MILD,1,1,1,0,0,0,1,0,1,1,0,0,1,1,0,0,0,0,2
A & N Islands,2007,MARCH,WEDNESDAY,MILD,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,MARCH,THURSDAY,MODERATE,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
A & N Islands,2007,MARCH,FRIDAY,MODERATE,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

Figura 15. Archivo de Extracción Origen

Talend Open Studio

Siguiendo con la implementación tenemos el proceso ETL (extraer, transformar, cargar) el cual es muy relevante dentro del análisis de datos. Existe una gran variedad de herramientas y programas que lo implementan para llevar a cabo el procesamiento de los macrodatos, por ejemplo, Talend Open Studio. Con la ayuda de esta herramienta creamos Jobs para poder realizar la extracción de los datos al csv Origen, generar archivos csv de las tablas que vamos a utilizar en el diseño de nuestra solución, transformar los datos y finalizar con la carga de los datos en Amazon S3. Los pasos por seguir para llevar a cabo la configuración de la herramienta son los siguientes:

1. Luego de haber instalado la herramienta en el equipo procedemos a ejecutar la aplicación y nos aparece la pantalla de inicio, Podemos seleccionar uno de los proyectos que tengamos creados en el nuestro espacio de trabajo. Si es la primera vez que entramos creamos un nuevo proyecto con "Create a new Project", también desde esta pantalla inicial, disponemos de la opción" Import an existing Project", que permite importar un proyecto a partir de un fichero local. Nosotros vamos a seleccionar la opción "Select an existing poject", seleccionar el nombre del proyecto y dar clic en el botón "Finish". la Imagen, muestra la selección de las opciones para poder abrir el proyecto correctamente.

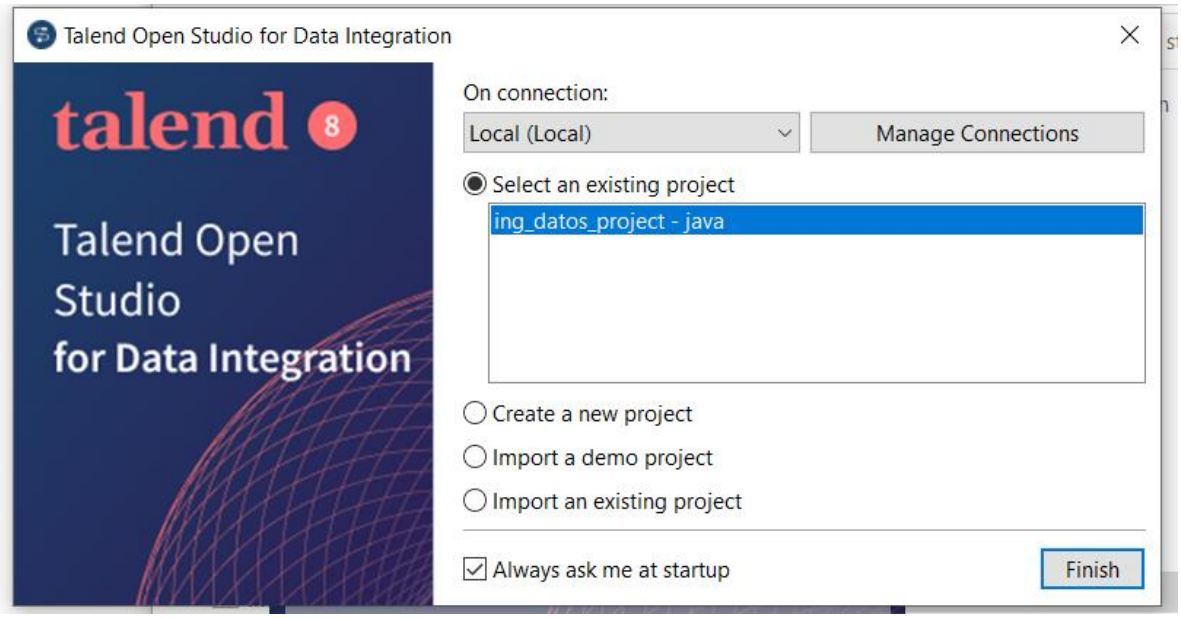


Figura 15. Pantalla de inicio de la aplicación

2. Crear la estructura de las carpetas para llevar un orden de trabajo, se crean las carpetas RAW, PROCESS y ACCESS en las cuales vamos a almacenar los diferentes Jobs creados para la extracción, carga y transformación de los datos (ETL). La estructura de las carpetas creadas en el proyecto se puede visualizar en la Imagen

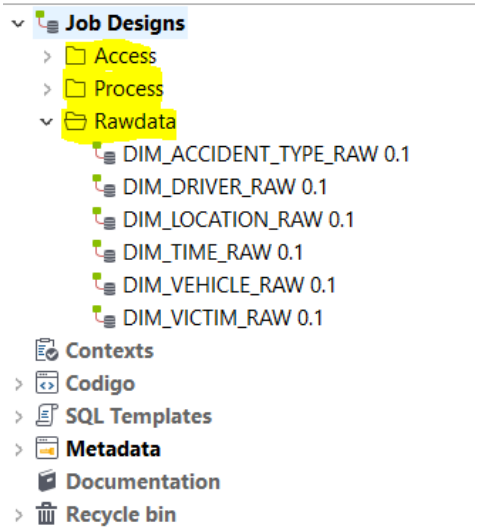


Figura 16. Estructura de carpetas del proyecto (RAW, PROCESS y ACCESS)

3. Creación del File Delimited de extracción de Origen y las carpetas para los pasos de RAW, PROCESS y ACCESS, nos ubicamos en la parte de "METADATA" y seleccionamos con click derecha la opción "FILE DELIMITED" y seleccionamos la opción de "CREATE FOLDER", creando los tres que necesitamos

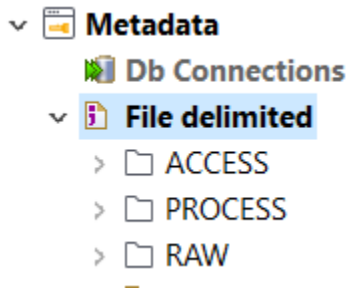


Figura 17. Carpetas Creadas en File Delimited

Luego debemos crear el archivo de extracción Origen, este paso se usará para todos los Jobs, crear el archivo delimitado para poder extraer la información, debido a que en todos los Jobs lo que se va generar es un archivo extensión csv

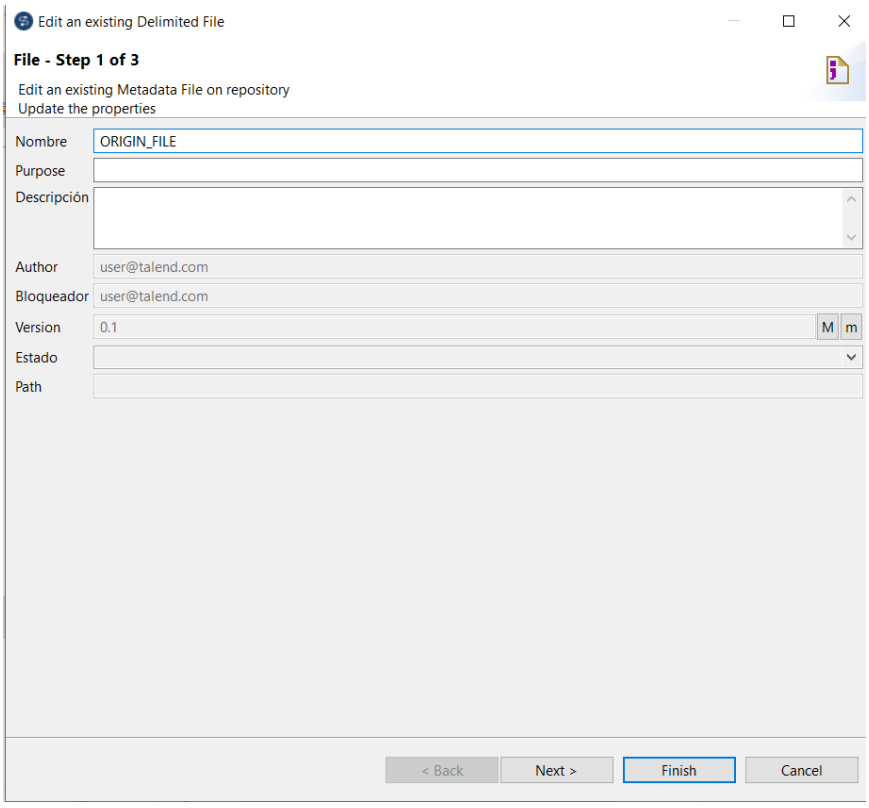


Figura 18. Configuración de Nombre de Archivo de extracción de csv

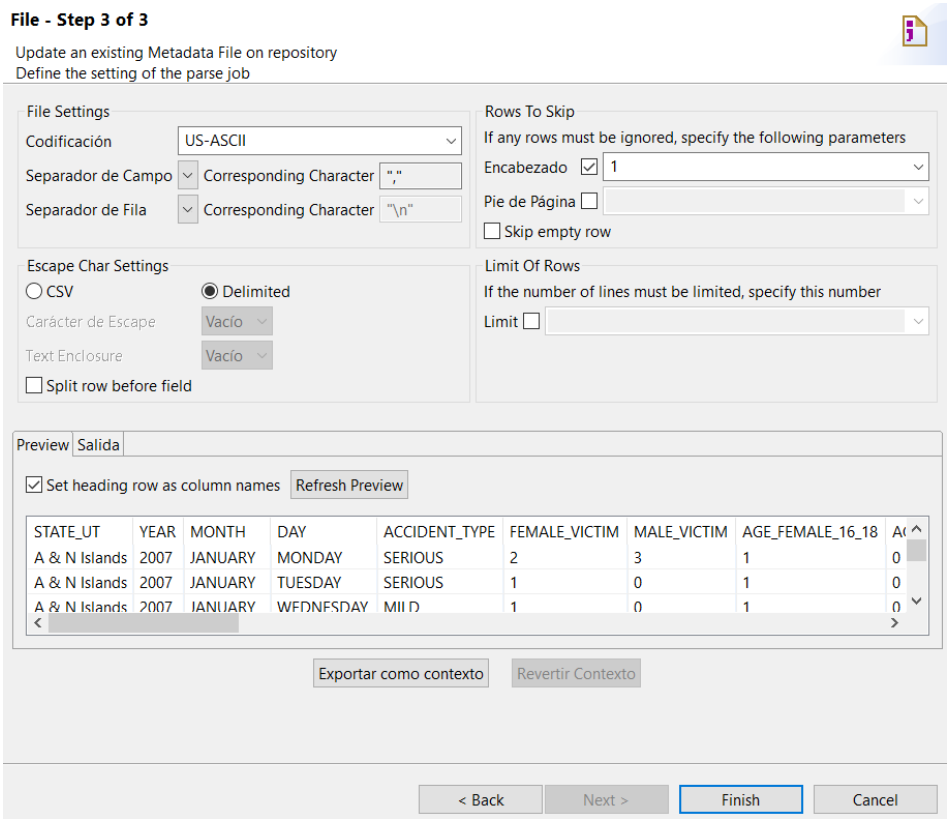


Figura 19. Previsualización de Archivo csv

4. Creación de los diferentes Jobs a utilizar en el ETL, vamos a dirigirnos a la carpeta creada y dar clic derecho para poder seleccionar la opción “Create Job”, luego de abrir la ventana para el diseño del Job, tenemos una paleta de opciones donde buscaremos aquellos componentes necesarios para la ejecución del ETL

- TFileInputDelimited: Este componente lee un archivo delimitado fila por fila para dividirlo en campos y luego envía los campos como se define en el esquema al siguiente componente.

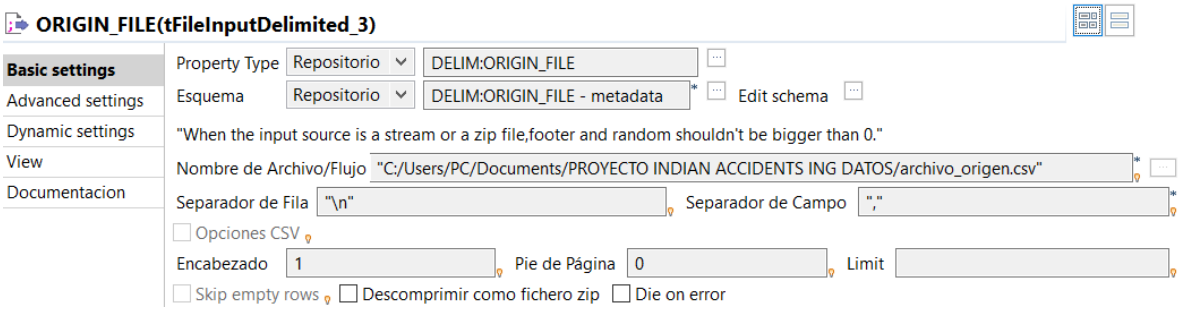


Figura 20. Configuración componente TFileInputDelimited

- Tmap: Este componente nos ayuda para poder manejar múltiples entradas y salidas, realizar uniones, transformaciones y más, en su configuración tendremos los datos de entrada con los datos de salida seleccionando los campos que vamos a necesitar mostrar en nuestra salida. La configuración de este componente la podemos visualizar.

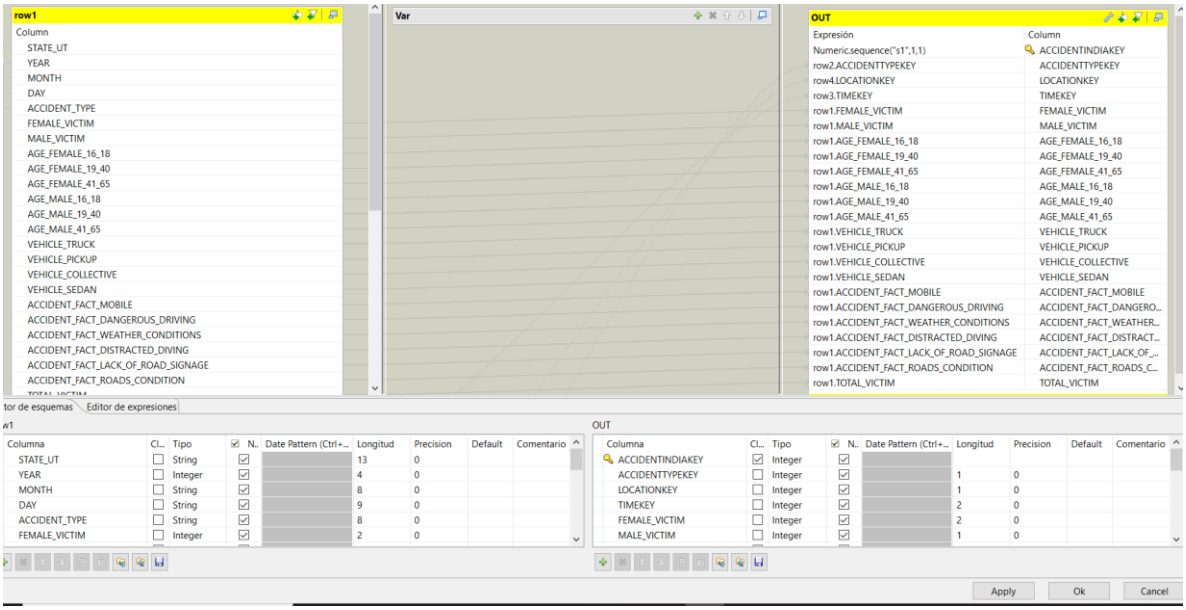


Figura 20. Editor Tmap

- tFileOutputDelimited: Este componente lee un archivo delimitado fila por fila para dividirlo en campos y luego envía los campos como se define en el esquema al siguiente componente. Esto quiere decir que genera un archivo csv y se almacena localmente dependiendo la dirección que le agreguen al asignar el nombre del archivo, y posteriormente se envía este archivo plano a nuestro almacenamiento en AWS S3.

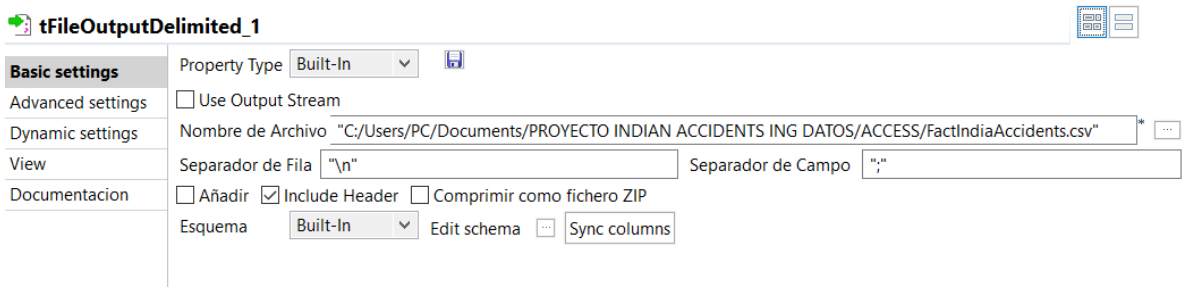


Figura 21. Configuración componente tFileOutputDelimited

- tS3Connection: Este componente permite establecer una conexión con Amazon S3 para almacenar y recuperar datos. El componente Standard tS3Connection pertenece a la familia Cloud. Se utilizará para poder realizar una conexión entre Talend Open Studio con AWS S3.

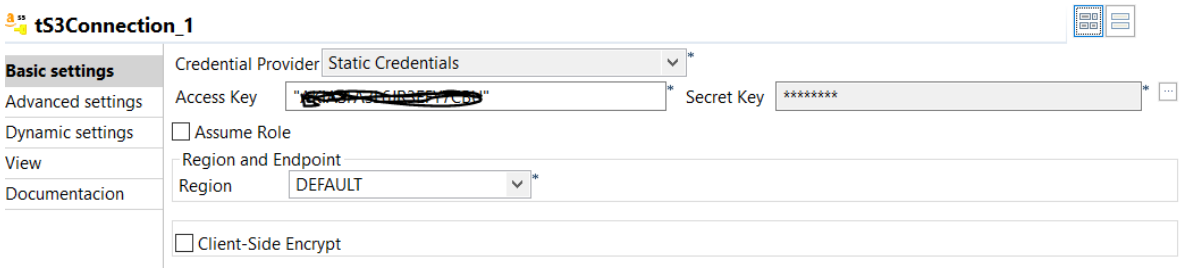


Figura 22. Configuración componente tS3Connection

- tS3Put: Carga datos en Amazon S3 desde un archivo local o desde la memoria caché a través del modo de transmisión. Para poder hacer un uso correcto de este componente necesitamos tener una conexión existente a AWS S3, se selecciona esa casilla para utilizar la conexión creada en el componente tS3Connection, se define el nombre del Bucket creado en S3 y la dirección en donde se almacenará el archivo csv que tenemos almacenado localmente en nuestra computadora.

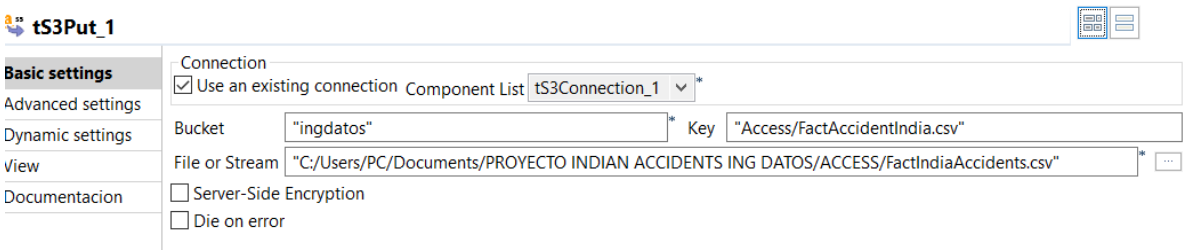


Figura 23. Configuración componente tS3Put

La siguiente imagen, refleja un ETL, el cual realiza el proceso lectura del archivo csv, extracción de los datos y guardado en un archivo de plano (CSV), la conexión con AWS S3 y finalizando con la carga del archivo de la tabla Time a la carpeta ACCESS del Bucket "ingdatos" creado en Amazon S3.

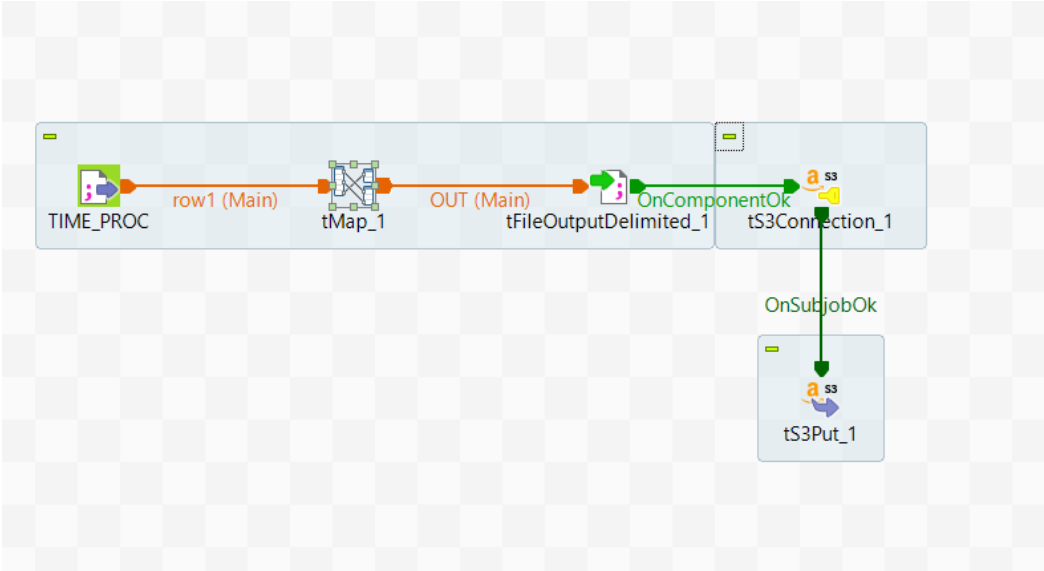


Figura 24. ETL para obtener un archivo csv y cargarlo en el bucket de Amazon S3.

Amazon S3

Se creó un bucket llamado “ingdatos” en Amazon S3 para la implementación de la solución.

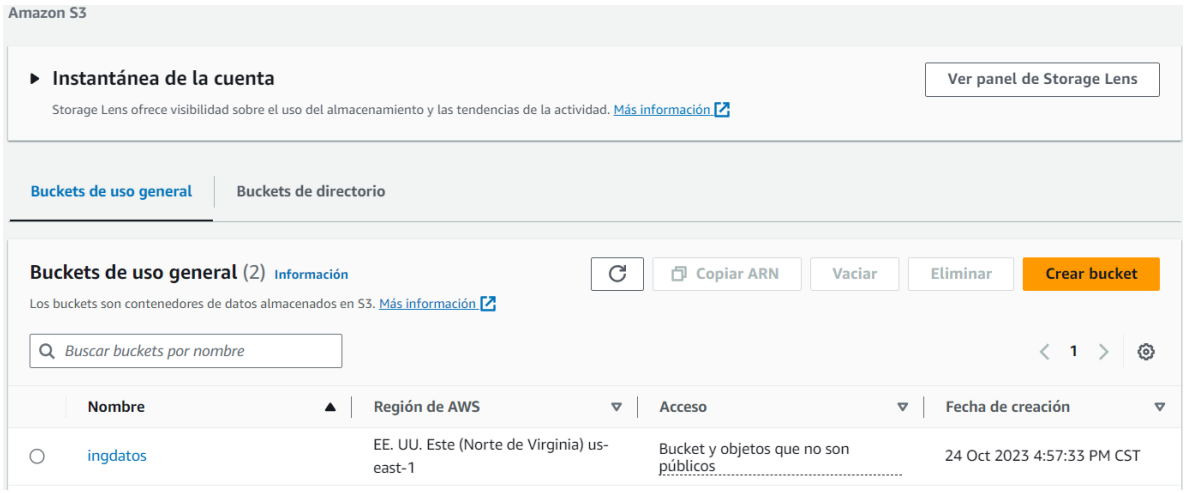


Figura 25. Bucket ingdatos, creado en Amazon S3

Creación de la estructura de almacenamiento en S3 con las siguientes carpetas dentro del bucket:

1. RAW.
2. PROCESS.
3. ACCESS.

La estructura de las carpetas creadas en Amazon S3 para el almacenamiento en la nube del proyecto.

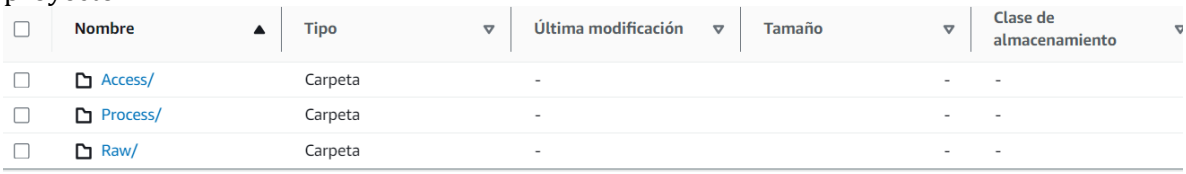


Figura 26. Estructura de almacenamiento en el Bucket de S3, carpetas RAW, PROCESS y ACCESS.

Para poder acceder al Bucket de Amazon S3 es necesario realizar una configuración al componente tS3Connection en Talend Open Studio, agregando el Access Key y la Secret Key para lograr el acceso a nuestro bucket “ingdatos”.

Amazon Redshift

La herramienta de Amazon Redshift se define como un servicio que se encarga de la gestión de procesos de operación, escalado y ajuste del almacenamiento de datos. La opción de Amazon Redshift se utilizará para obtener un almacenamiento de la base de datos en la nube, para ello se llevó a cabo la creación de un clúster en AWS Redshift con el nombre “uescluster1”, en el cual vamos a poder almacenar los datos de la capa de presentación obtenidos del proceso ETL de Talend Open Studio y su carga de datos en Amazon S3 para su posterior visualización gráfica en Power BI y ayudar a la toma de decisiones estratégicas.

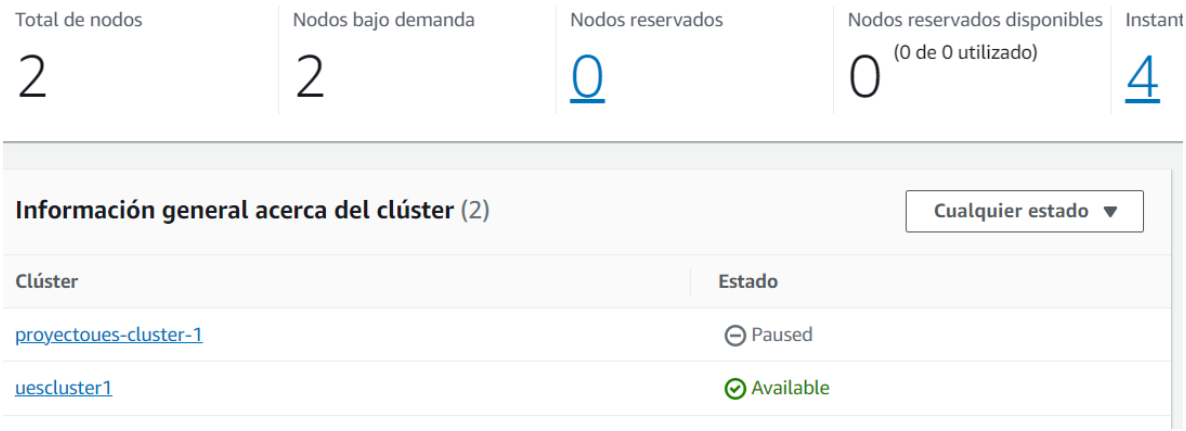


Figura 27. Cluster uescluster1, creado en Amazon Redshift

Para lograr la conexión entre Redshift y la capa de PRESENTATION de S3 se ejecutó el siguiente procedimiento:

Ejecutamos un script con el cual creamos la estructura de la tabla de hecho y sus dimensiones. Como ejemplo: Dimensión Locación

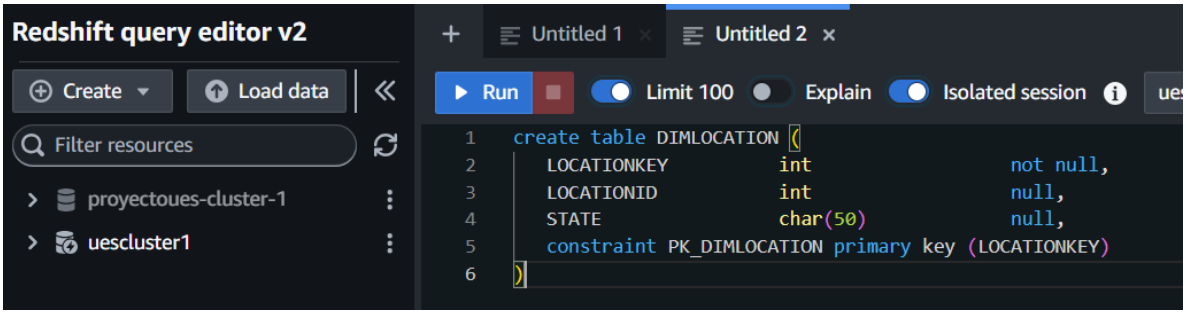


Figura 27. Script de creación de estructura de Dimensión Locación

Luego, se carga los datos de la Dimensión Locación que se encuentra en la capa PRESENTATION de S3.

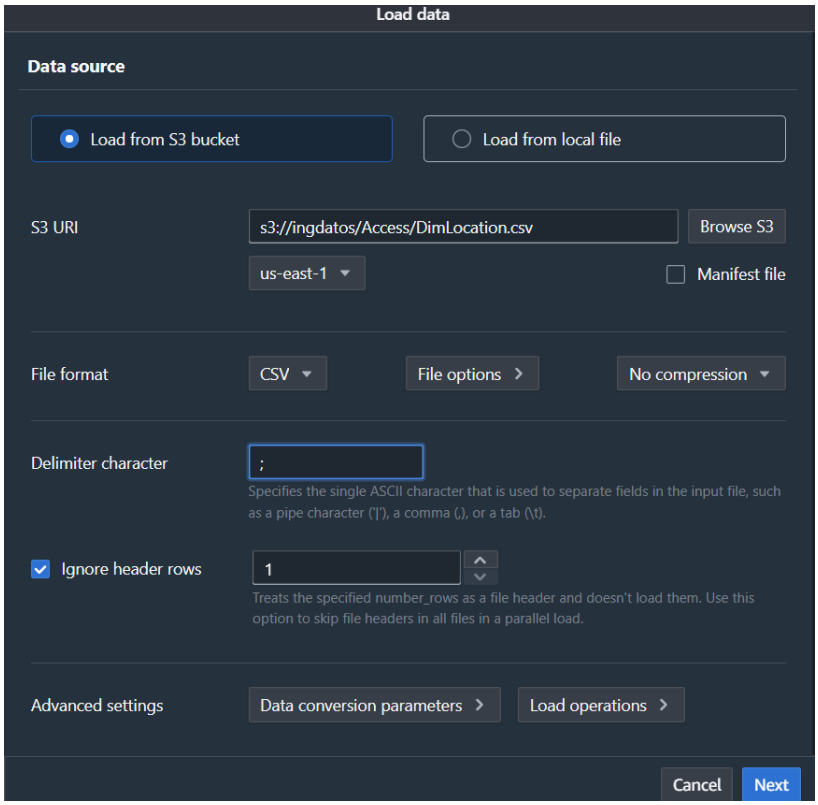


Figura 28. Selección del archivo del bucket a cargar en la tabla de RedShift

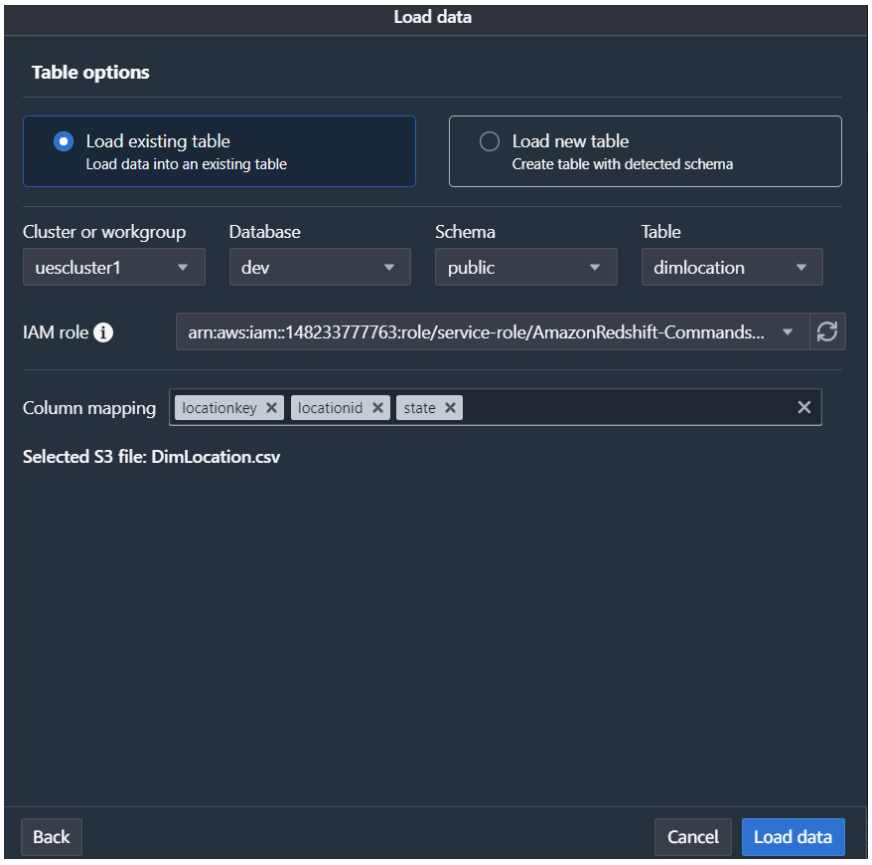


Figura 29. Selección del esquema y tabla donde se cargarán los datos

Power BI

Finalmente se utilizará la herramienta Power BI ya que es una solución de análisis empresarial basado en la nube, y permite unir diferentes fuentes de datos, analizarlos y presentar un análisis de estos a través de informes y paneles. Realizando una conexión con la base de datos de AWS Redshift podemos acceder a los datos almacenados en la base de datos con el modelo del DataWarehouse definido en la implementación de la solución, obtenido del ETL generado con Talend Open Studio y almacenado en la capa de presentación en AWS S3. Estos análisis pueden ser compartidos por diferentes usuarios de la misma organización; por lo que directivos, financieros, comerciales, etc., pueden disponer de la información del negocio en tiempo real y poder tomar decisiones estratégicas con la información obtenida. En la Imagen 26, podemos visualizar el primer paso para seleccionar la base de datos de Amazon Redshift en Power BI para poder realizar la conexión. La Imagen, muestra los campos obligatorios para poder conectar Power BI con Amazon Redshift. Muestra las tablas de la base de datos en Redshift luego de la conexión.

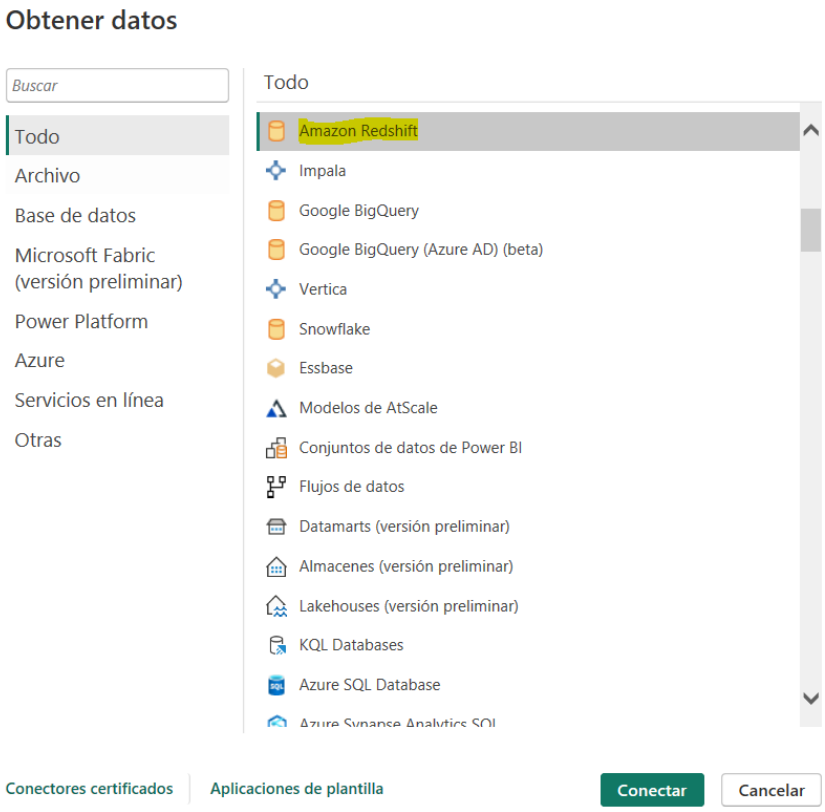
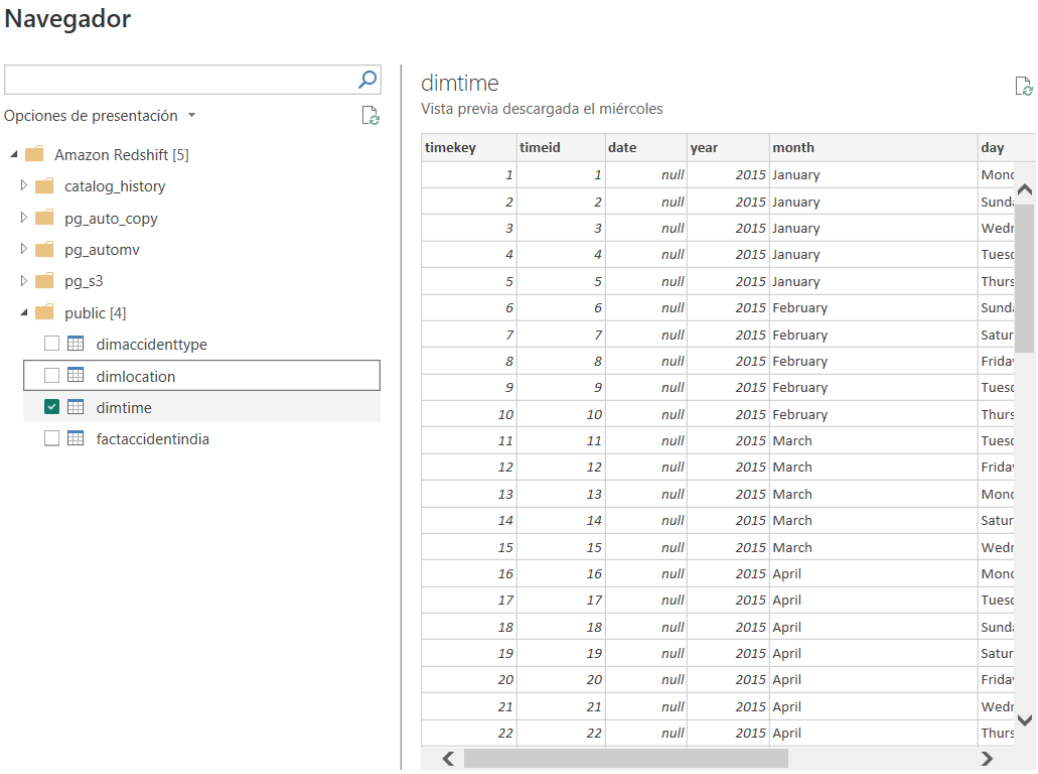


Figura 30. Conexión Power BI con Redshift

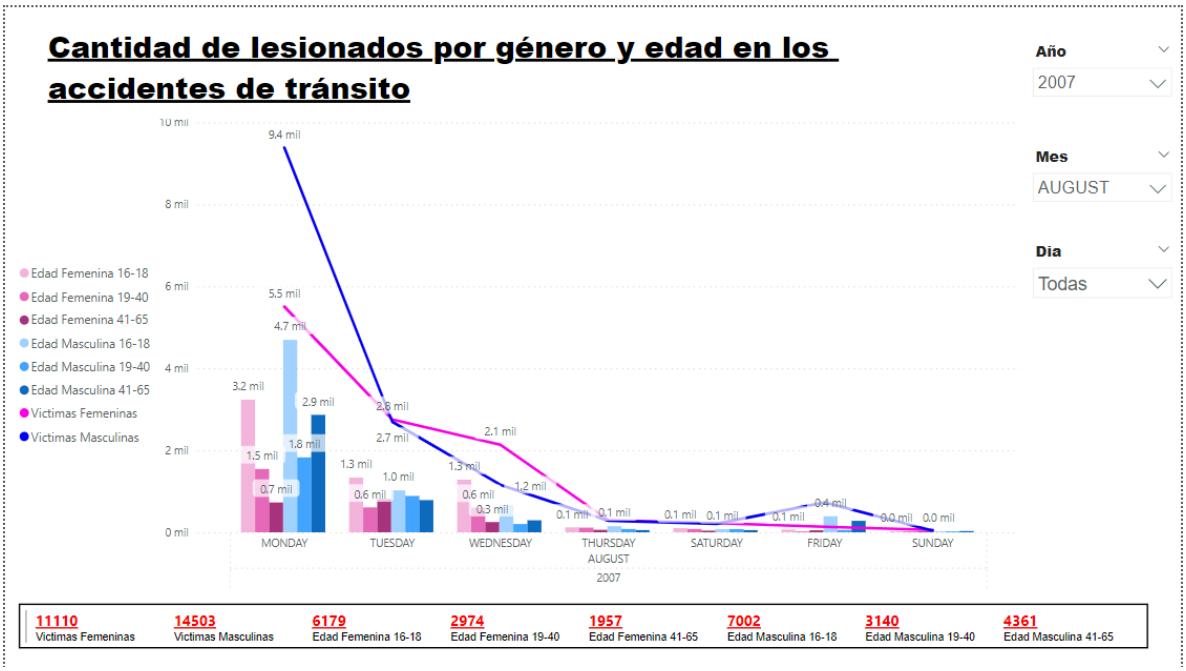
Figura 31. Conexión con el clúster de RedShift



timekey	timeid	date	year	month	day
1	1	null	2015	January	Mond
2	2	null	2015	January	Sund
3	3	null	2015	January	Wedn
4	4	null	2015	January	Tuesd
5	5	null	2015	January	Thurs
6	6	null	2015	February	Sund
7	7	null	2015	February	Satur
8	8	null	2015	February	Friday
9	9	null	2015	February	Tuesd
10	10	null	2015	February	Thurs
11	11	null	2015	March	Tuesd
12	12	null	2015	March	Friday
13	13	null	2015	March	Mond
14	14	null	2015	March	Satur
15	15	null	2015	March	Wedn
16	16	null	2015	April	Mond
17	17	null	2015	April	Tuesd
18	18	null	2015	April	Sund
19	19	null	2015	April	Satur
20	20	null	2015	April	Friday
21	21	null	2015	April	Wedn
22	22	null	2015	April	Thurs

Figura 32. Carga de Tabla de Hechos y Dimensiones

La implementación de un Data Warehouse proporciona una visión más clara y detallada del negocio, lo que permite a los altos mandos tomar las mejores decisiones en el menor tiempo posible.



b) Figura 33. Dasboard en PoweBI

c) Presupuesto de implementación

Al ser un proyecto nuevo de ser solicitado por una empresa de seguros viales para poder evaluar los tipos de planes a ofrecer; por tal razón es de suma importancia contratar servicios profesionales para la implementación de la solución, y posteriormente capacitaciones a los recursos de la empresa en el uso correcto de las herramientas necesarias para la visualización de los informes estratégicos que ayuden en la toma de decisiones. Dicho costo se realizará tomando en cuenta la implementación y capacitación que se llevará a cabo en un lapso de 10 días teniendo un costo total de $7500 en esto considerando salario de los dos especialistas más las capacitaciones al equipo de IT

.

Hardware

Es necesario que la empresa cuente con un equipo que cumpla con los recursos mínimos y que sea capaz de soportar la ejecución simultanea de las diversas herramientas utilizadas en el desarrollo de la solución. Tomando en cuenta esto se optó por la compra de un equipo con las siguientes características:

- Procesador i9-12900K
- Disco SSD 1 TB
- Memoria RAM 16 GB (como mínimo)
- Velocidad de Internet 50 MB

Generando un costo total de $ 800.00 siendo un total la inversión completa de $ 8300.00

d) Análisis de resultados

Especificación de Necesidades que el Modelo Propuesto resolverá.
Reporte de total de accidentes de tránsito: Representa el total de víctimas que han tenido accidentes de tránsito de acuerdo al año, mes y día.

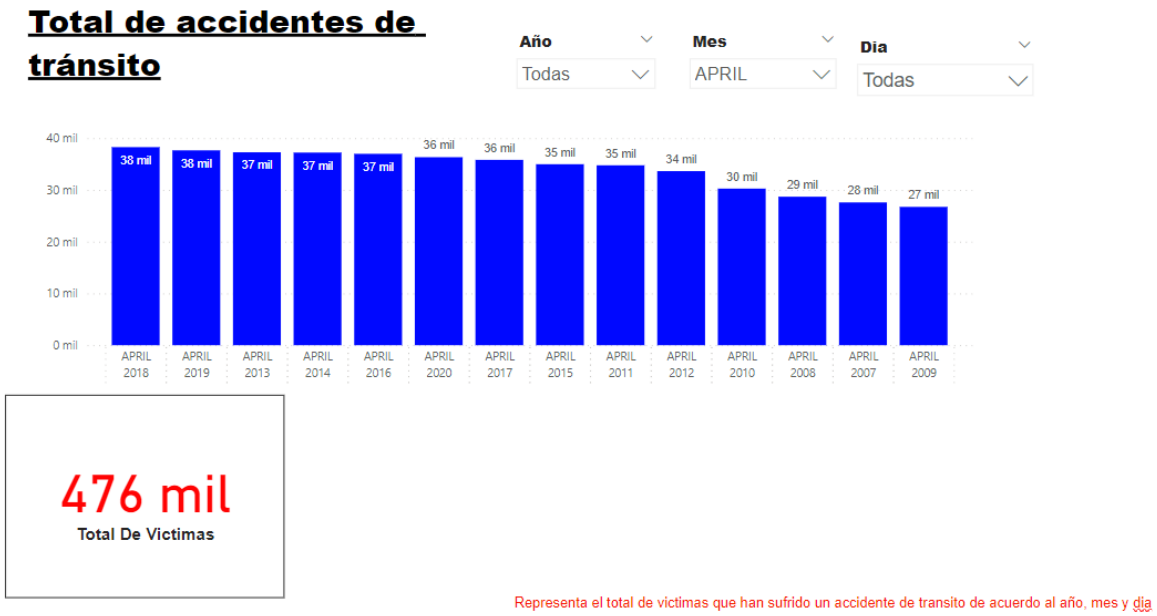


Figura 33 *Total de accidentes de transito*

Reporte de la cantidad de lesionados por género y edad en los accidentes de tránsito: Representa la cantidad de víctimas en accidentes de tránsito de acuerdo a su género y rango de edades de acuerdo al año, mes y día.

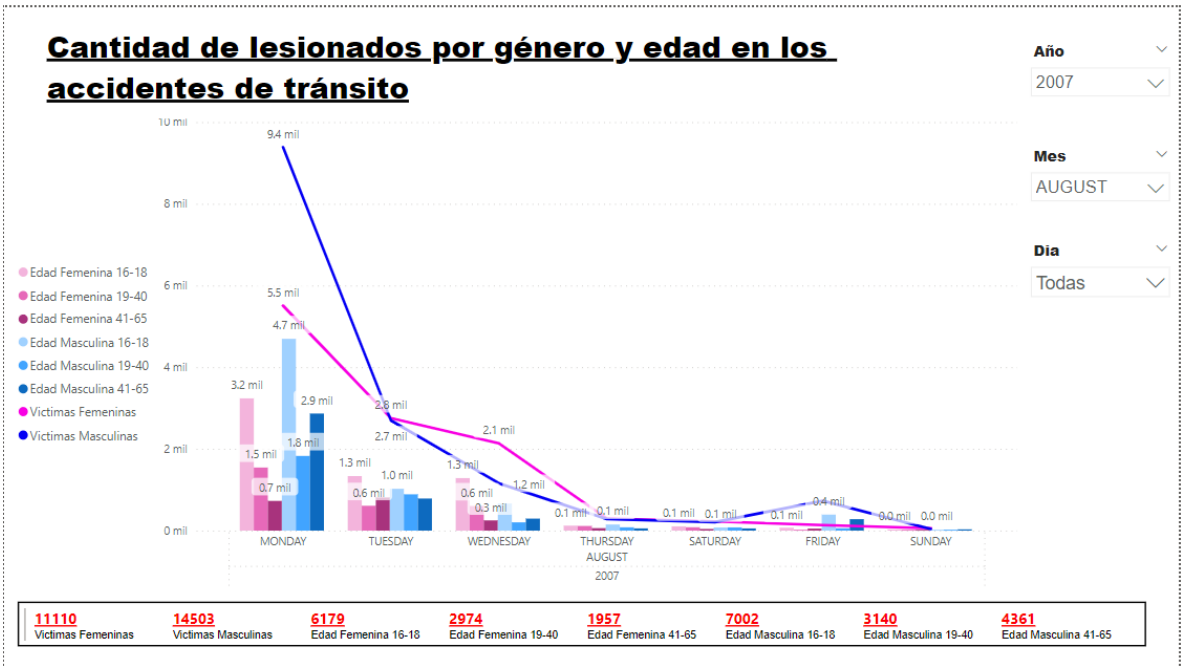


Figura 34 *Cantidad de lesionados por género y edad en los accidentes de transito*

Reporte de cantidad de accidentes por ubicación: Representa la cantidad de víctimas en accidentes de tránsito de acuerdo a la ciudad, año, mes y día. Teniendo una mejor perspectiva de que ciudad es la que frecuenta más accidentes de tránsito.

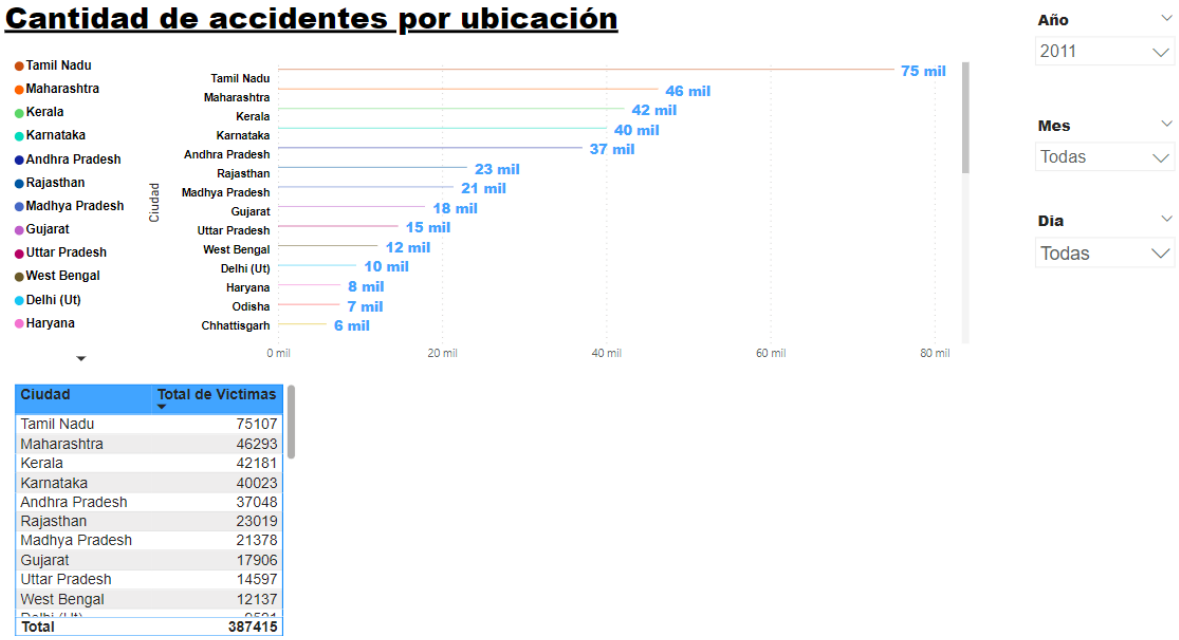


Figura 35 Cantidad de accidentes por ubicación

Reporte de total de accidentes por tipo de vehículo: Podemos visualizar que tipo de vehículo es el que más frecuenta accidentes de tránsito de acuerdo al año, mes y día.

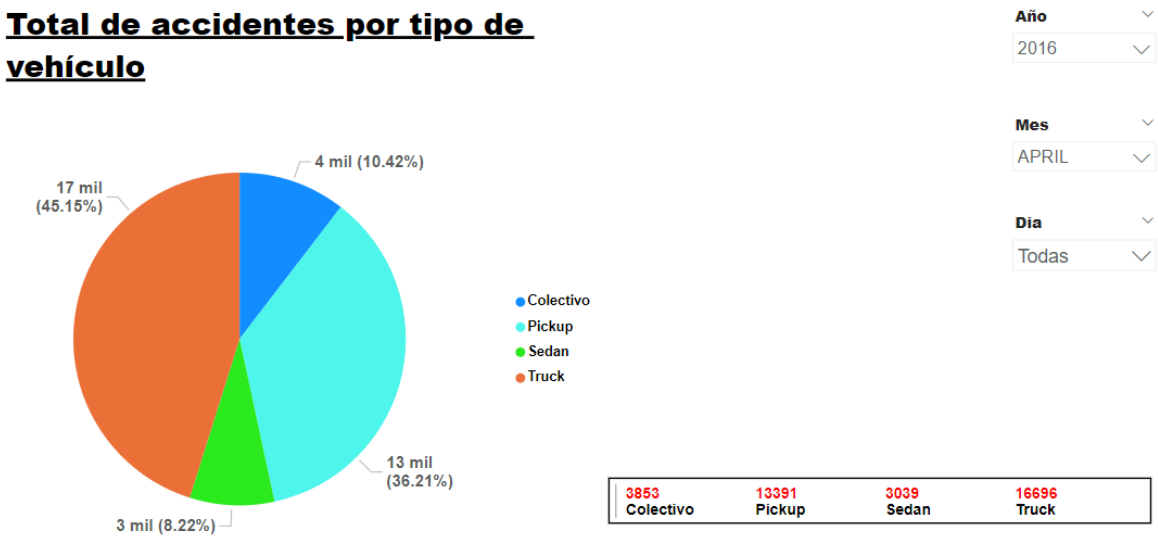


Figura 36 Total de accidentes por tipo de vehículo

Reporte de total de víctimas por tipo de accidente: En el siguiente dashboard podemos visualizar por qué motivo las personas tienden a tener accidentes de tránsito y poder tener alguna contra medida hacia los conductores que manejan irresponsablemente.

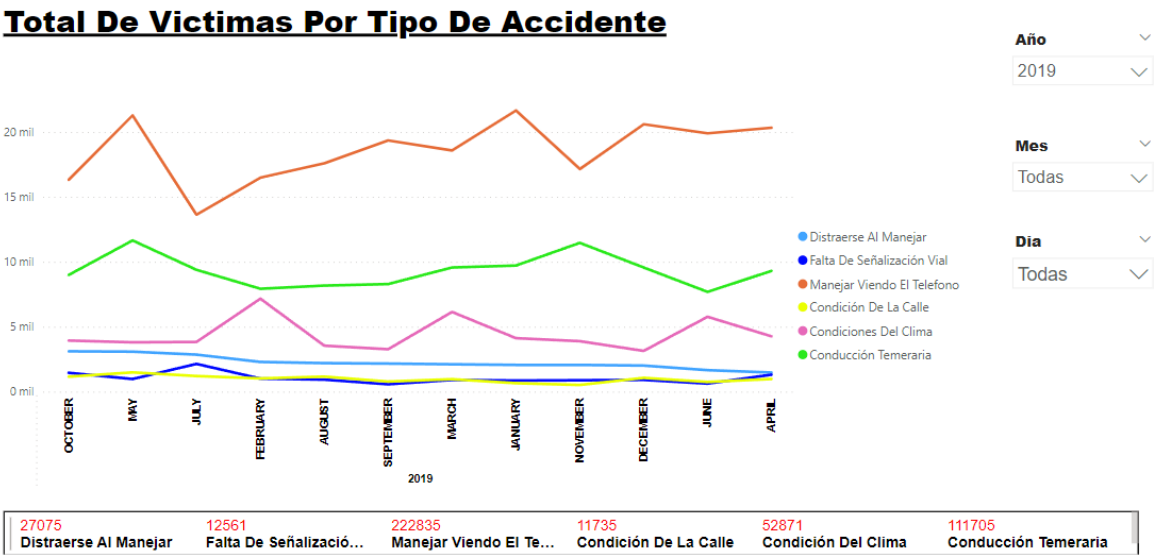


Figura 37 total de víctimas por tipo de accidente

5) Conclusiones y Recomendaciones

Concluimos que los procesos ETL son la parte más fundamental del diseño de la solución, ya que, de cierta manera, son el puente entre los grandes volúmenes de datos de los csv, con los cuales resulta imposible analizarlos y tomar decisiones con base en ellos; y finalizan en la herramienta de Business Intelligence, la cual cumple el objetivo de apoyar en la toma de decisiones de la parte gerencial de la empresa. Todo esto es posible mediante el proceso ETL que sufren los datos desde su concepción hasta su culminación en la herramienta de BI.

El desarrollo una solución de Big data, le permitirá contar con un Data Werehouse, que de manera directa dará solución a los principales requerimientos analíticos, que condujeron al desarrollo del proyecto. Es importante mencionar que los datos finales además podrán ser aplicados en soluciones de ciencia de datos, aplicación de Machine Learning, Inteligencias de Negocio, es decir que el modelo final está dotado de la capacidad para poder alimentar procesos superiores de análisis de datos, incrementando así el potencial de la solución.

La creación de reportes que dan solución concreta a los requerimientos de negocio, permite entender realmente el enorme aporte analítico, que los datos procesados le puedan dar a las organizaciones, las innumerables inquietudes, preguntas o necesidades que pueden surgir en el tiempo. Ya que un reporte permite transmitir de manera directa los hallazgos encontrados en los datos, hallazgos que pueden incrementar el éxito de la toma de decisiones estratégica.

Debido al gran volumen de datos generados en los csv, es primordial contar con una potente herramienta de almacenamiento (S3), con gran disponibilidad, seguridad y soporte todos los procesos de ETL de la solución. De la misma manera se debe contar con una herramienta que soporte consultas de análisis de datos (Redshift), con un alto un rendimiento a través de las mismas herramientas basadas en SQL y aplicaciones de inteligencia empresarial que la solución implemente.

6) Bibliografía

- https://www.kaggle.com/

- Kimball, R., & Ross, M. (2011). The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons.

- Ramos, S. (2016). Data Warehouse, data marts y modelos dimensionales. Un pilar fundamental para la toma de decisiones. Albatera: SolidQ.

- Kimball. (2013). Business Intelligence Applications. In R. &. Kimball, The data warehouse toolkit (3rd ed.) (pp. 22-26). Indianapolis: Wiley.

- Kimball. (2013). Dimension Tables for Descriptive Context. In R. &. Kimball, The data warehouse toolkit (3rd ed.) (pp. 13-15). Indianapolis: Wiley.

- Kimball. (2013). Dimensional Modeling Introduction. In R. &. Kimball, The data warehouse toolkit (3rd ed.) (pp. 7-8). Indianapolis: Wiley.

- Kimball. (2013). Extract, Transformation, and Load System. In R. &. Kimball, The data warehouse toolkit (3rd ed.) (pp. 19-21). Indianapolis: Wiley.

- Kimball. (2013). Fact Tables for Measurements. In R. &. Kimball, The data warehouse toolkit (3rd ed.) (pp. 10-12). Indianapolis: Wiley.

- Kimball. (2013). Facts and Dimensions Joined in a Star Schema. In R. &. Kimball, The data warehouse toolkit (3rd ed.) (pp. 16-18). Indianapolis: Wiley.

- Kimball. (2013). Kimball's DW/BI Architecture. In R. &. Kimball, The data warehouse toolkit (3rd ed.) (pp. 18-22). Indianapolis: Wiley.