Sistema de Recomendación de Libros de Biblioteca

Mamani Ayala, Brandon (2015052715), Quispe Mamani, Angelo (2015052826), Vizcarra Llanque, Jhordy (2015052719), Ordoñez Quilli, Ronald (2015052821), Rodriguez Mamani, Juan (2017057862)

Tacna, Perú

Abstract

The recommendation systems are part of an information filtering system, which present different types of topics or information items (movies, music, books, news, images, web pages, etc.) that are of interest to a user. particular. In this short article we will try to provide a general explanation about the different data models, such as Content Filtering and Collaborative Filtering. Then we will apply these models in an example with the help of Colab, where we will filter data and show the results as well as their different graphs. Then we will give some appreciations and conclusions of what was our Final Work.

1. Resumen

Los sistemas de recomendación forman parte de un sistema de filtrado de información, los cuales presentan distintos tipos de temas o ítems de información (películas, música, libros, noticias, imágenes, páginas web, etc.) que son del interés de un usuario en particular. En este breve artículo intentaremos brindar una explicación general acerca de los diferentes modelos de datos, tales como el Filtrado por Contenido y Filtrado Colaborativo. Luego se aplicarán dichos modelos en un ejemplo con ayuda de Colab, donde filtraremos datos y mostraremos los resultados así como sus diferentes gráficas. Luego daremos unas apreciaciones y conclusiones de lo que fue nuestro Trabajo Final.

2. Introduccion

Todos hemos estado expuestos a los famosos "filtros colaborativos": cuando estabas comprando online y automáticamente el sitio web te recomienda otro artículo similar, o cuando le diste "me gusta" a algo y de repente te empezaron a salir recomendaciones similares en una red social. Los sistemas de recomendación evalúan patrones de tu comportamiento y de miles de usuarios a la vez para emitir nuevas recomendaciones. Ésta es una de las aplicaciones más comunes de "Machine Learning", porque te da la sensación de estar navegando en un sitio web programado únicamente para tí.

La mayoría de sistemas de recomendación de productos usan filtros colaborativos. Estamos en una era donde nosotros como usuarios generamos más información que nunca antes en la historia. Esta información es usada por los servicios que usamos a diario para darnos una experiencia mas personalizada.

Un claro ejemplo de esto es Netflix, ellos usan la información que nosotros generamos para hacernos recomendaciones y mostrarnos categorías y películas según nuestros gustos. Otro ejemplo es Amazon, que hace recomendaciones de productos que nos podrían interesar, basándose en los productos que ya compramos.

El filtro colaborativo es una técnica usada por los sistemas de recomendación, se basa en el hecho de que si dos personas X y Y en un mismo sistema tienen gustos similares, entonces recomendarle a la persona X cosas que le gusten a la persona Y será de gran relevancia, en contraste a simplemente darle una opción cualquiera.

Hay una variación de filtro colaborativo que se basa en comparar los items en vez de los usuarios para arrojar los items relacionados, de modo que si un usuario entra a ver un determinado item, le podemos recomendar otro.

3. Marco Teorico

3.1 Introduccion.

Un almacen de datos (data warehouse, DW) segun Inmon (Inmon 02, Imhoff y Galemmo 03), es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se trata, sobre todo, de un historial completo de la organización, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos (especialmente con herramientas OLAP, de procesamiento analítico en línea). Por otra parte Kimball (Kimball 98) la define como "una copia de los datos transaccionales estructurados específicamente para consultas y analisis". Actualmente uno de los mayores impedimentos para construir este tipo de almacenes de datos es la falta de conocimiento de metodologías adecuadas para su implementación, y la disciplina para cumplirlas. En este breve artículo describiremos la metodología más utilizada actualmente: la metodología de Kimball.

3.2 Metodologias Actuales

Existen muchas metodologías de diseño y construcción de DW. Cada fabricante de software de inteligencia de negocios busca imponer una metodología con sus productos. Sin embargo, se imponen entre la mayoría dos metodologías, la de Kimball y la de Inmon. Para comprender la mayor diferencia entre estas dos metodologías, debemos explicar además de la noción de DW mencionando en la introducción, la idea de Data mart. Un Data mart (Kimball et al 98) es un repositorio de información, similar a un DW, pero orientado a un área o departamento específico de la organización (por ejemplo Compras, Ventas, RRHH, etc.), a diferencia del DW que cubre toda la organización, es decir la diferencia fundamental es su alcance. Desde el punto de vista arquitectónico, la mayor diferencia entre los dos autores es el sentido de la construcción del DW, esto es comenzando por los Data marts o ascendente (Bottom-up, Kimball) o comenzando con todo el DW desde el principio, o descendente. Por otra parte, la metodología de Inmon se basa en conceptos bien conocidos del diseño de bases de datos relacionales (Inmon 02, Imhoff y

Galemmo 03) la metodología para la construcción de un sistema de este tipo es la habitual para construir un sistema de información, utilizando las herramientas habituales, al contrario de la de Kimball, que se basa en un modelado dimensional (no normalizado)

3.3 Metodologia Inmon

Bill Inmon ve la necesidad de transferir la información de los diferentes OLTP (Sistemas Transaccionales) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el analisis (sería el CIF o Corporate Information Factory).
Insiste ademas en que ha de tener las siguientes características:

- Orientado a temas: Los datos en la base de datos están estructurados de forma que los datos estén relacionados.
- Integrado: La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.
- No volátil: La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.
- Variante en el tiempo: Los datos que sufren un cambio a través del tiempo, deben ser registrados para que los informes generados reflejen esas variaciones.

La información ha de estar a los máximos niveles de detalle. Los Dw departamentales o datamarts son tratados como subconjuntos de este Dw corporativo, que son construidos para cubrir las necesidades individuales de analisis de cada departamento, y siempre a partir de este Dw Central (del que también se pueden construir los ODS (Operational Data Stores) o similares).

El enfoque Inmon tambien se referencia normalmente como Top-down. Los datos son extraidos de los sistemas operacionales por los procesos ETL y cargados en las areas de stage, donde son validados y consolidados en el DW corporativo, donde ademas existen los llamados metadatos que documentan de una forma clara y precisa el contenido del DW. Una

vez realizado este proceso, los procesos de refresco de los Data Mart departamentales obtienen la información de el, y con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos, refrescando su contenido.

- Definición de Data Warehouse: Un Data Warehouse proporciona una visión global, común e integrada de los datos de la organización, independiente de cómo se vayan a utilizar posteriormente por los consumidores o usuarios. Normalmente en el almacén de datos habrá que guardar información histórica que cubra un amplio período de tiempo.

- Definición de Data Mart: Podemos entender un Data Mart como un subconjunto de los datos del Data Warehouse con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica. Al igual que en un data warehouse, los datos están estructurados en modelos de estrella o copo de nieve y un data mart puede ser dependiente o independiente de un data warehouse.

3.4 Metodologia Kimball en detalle.

La metodología se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle) (Kimball et al 98, 08, Mundy y Thornthwaite 06). Este ciclo de vida del proyecto de DW, está basado en cuatro principios básicos :

- Centrarse en el negocio: Hay que concentrarse en la identificación de los requerimientos del negocio y su valor asociado, y usar estos esfuerzos para desarrollar relaciones sólidas con el negocio, agudizando el análisis del mismo y la competencia consultiva de los implementadores.

- Construir una infraestructura de información adecuada: Diseña una base de información única, integrada, fácil de usar, de alto rendimiento donde se reflejará la amplia gama de requerimientos de negocio identificados en la empresa.

- Realizar entregas en incrementos significativos crear el almacén de datos (DW) en incrementos entregables en plazos de 6 a 12 meses. Hay que usa el valor de negocio de cada element identificado para determinar el orden de aplicación de los incrementos. En esto la metodología se parece a las metodologías ágiles de construcción de software.

- Ofrecer la solución completa proporcionar todos los elementos necesarios para entregar valor a los usuarios de negocios. Para comenzar, esto significa tener un almacén de datos sólido, bien diseñado, con calidad probada, y accesible. También se deberá entregar herramientas de consulta ad hoc, aplicaciones para informes y análisis.

La construcción de una solución de DW/BI (Datawarehouse/Business Intelligence) es sumamente compleja, y Kimball nos propone una metodología que nos ayuda a simplificar esa complejidad. Las tareas de esta metodología (ciclo de vida) se muestran en la figura 1. De la figura 1, podemos observar dos cuestiones. Primero, hay que resaltar el rol central de la tarea de definición de requerimientos. Los requerimientos del negocio son el soporte inicial de las tareas subsiguientes. También tiene influencia en el plan de proyecto (nótese la doble fecha entre la caja de definición de requerimientos y la de planificación). En segundo lugar podemos ver tres rutas o caminos que se enfocan en tres diferentes áreas

De la figura 1, podemos observar dos cuestiones. Primero, hay que resaltar el rol central de la tarea de definición de requerimientos. Los requerimientos del negocio son el soporte inicial de las tareas subsiguientes. También tiene influencia en el plan de proyecto (nótese la doble fecha entre la caja de definición de requerimientos y la de planificación). En segundo lugar podemos ver tres rutas o caminos que se enfocan en tres diferentes áreas:

- Tecnología (Camino Superior). Implica tareas relacionadas con software específico, por ejemplo, Microsoft SQL Analysis Services.

- Datos (Camino del medio). En la misma diseñaremos e implementaremos el modelo dimensional, y desarrollaremos el subsistema de Extracción, Transformación y Carga (Extract, Transformation, and Load ETL) para cargar el DW.

- Aplicaciones de Inteligencia de Negocios (Camino Inferior). En esta ruta se encuentran tareas en las que diseñamos y desarrollamos las aplicaciones de negocios para los usuarios finales.

Estas rutas se combinan cuando se instala finalmente el sistema. En la parte de debajo de la figura se muestra la actividad general de administración del proyecto. A continuación describiremos cada una de las tareas.

4. Proceso de Desarrollo de Modelo Kimball

4.1 Planificacion del Proyecto

En este proceso se determina el propósito del proyecto de DW/BI, sus objetivos específicos y el alcance del mismo, los principales riesgos y una aproximación inicial a las necesidades de información.

Esta tarea incluye las siguientes acciones típicas de un plan de proyecto:

- Definir el alcance (entender los requerimientos del negocio).
- Identificar las tareas
- Programar las tareas
- Planificar el uso de los recursos.
- Asignar la carga de trabajo a los recursos
- Elaboración de un documento final que representa un plan del proyecto.
- Monitoreo del estado de los procesos y actividades.

- Rastreo de problemas
- Desarrollo de un plan de comunicación comprensiva que direccione la empresa y las áreas de TI

4.2 Definicion de Requisitos de Negocios

La definición de requerimientos, es un proceso de entrevistar al personal de negocio y técnico, aunque siempre conviene, tener un poco de preparación previa. En esta tarea, se debe aprender sobre el negocio, los competidores, la industria y los clientes del mismo. Se debe dar una revision a todos los informes posibles de la organización; rastrear los documentos de estrategia interna; entrevistar a los empleados, analizar lo que se dice en la prensa acerca de la organización, la competencia y la industria y se deben conocer los términos y la terminología del negocio.

4.3 Modelo Dimensional

Es un proceso dinámico y altamente iterativo. Comienza con un modelo dimensional de alto nivel obtenido a partir de los procesos priorizados y descritos en la tarea anterior, y El proceso iterativo consiste en cuatro pasos:

- Elegir el proceso de negocio
- Establecer el nivel de granularidad
- Elegir las dimensiones
- Identificar medidas y las tablas de hechos

4.4 Diseño Fisico

En esta tarea, se contestan las siguientes preguntas:

- ¿Cómo puede determinar cuán grande será el sistema de DW/BI?
- ¿Cuáles son los factores de uso que llevarán a una configuración más grande y más compleja?
- ¿Cómo se debe configurar el sistema?

- ¿Cuánta memoria y servidores se necesitan? ¿Qué tipo de almacenamiento y procesadores?
- ¿Cómo instalar el software en los servidores de desarrollo, prueba y producción?
- ¿Qué necesitan instalar los diferentes miembros del equipo de DW/BI en sus estaciones de trabajo?
- ¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?
- ¿Cómo conseguir un plan de indexación inicial?
- ¿Debe usarse la partición en las tablas relacionales?

4.5 Diseño e Implementacion de Subsistema de Extraccion, Transicion y Carga

El subsistema de Extracción, Transformación y Carga (ETL) es la base sobre la cual se alimenta el Data warehouse. Si se diseña adecuadamente, puede extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el DW en un formato acorde para la utilización por parte de las herramientas de análisis.

4.6 Mantenimiento y Crecimiento del Data Warehouse

Para administrar el entorno del Data Warehouse existente es importante enfocarse en los usuarios de negocio, los cuales son el motivo de su existencia, además de gestionar adecuadamente las operaciones del Data Warehouse, medir y proyectar su éxito y comunicarse constantemente con los usuarios para establecer un flujo de retroalimentación, En esto consiste el Mantenimiento. Finalmente, es importante sentar las bases para el crecimiento y evolución del Data Warehouse en donde el aspecto clave es manejar el crecimiento y evolución de forma iterativa utilizando el Ciclo de Vida propuesto, y establecer las oportunidades de crecimiento y evolución en orden por nivel prioridad.

4.7 Especificacion de Aplicaciones de BI

En está tarea se proporciona, a una gran comunidad de usuarios una forma más estructurada y por lo tanto, más fácil, de acceder al almacén de datos. Se proporciona este acceso estructurado a través de lo que llamamos, aplicaciones de inteligencia de negocios (Business Intelligence Aplications). Las aplicaciones de BI son la cara visible de la inteligencia de negocios: los informes y aplicaciones de análisis proporcionan información útil a los usuarios. Las aplicaciones de BI incluyen un amplio espectro de tipos de informes y herramientas de análisis, que van desde informes simples de formato fijo, a sofisticadas aplicaciones analíticas que usan complejos algoritmos e información del dominio. Kimball divide a estas aplicaciones en dos categorías basadas en el nivel de sofisticación, y les llama:

- Informes Estandar
- Aplicaciones Analiticas

4.8 Mantenimiento y Crecimiento del Data Warehouse

Para administrar el entorno del Data Warehouse existente es importante enfocarse en los usuarios de negocio, los cuales son el motivo de su existencia, además de gestionar adecuadamente las operaciones del Data Warehouse, medir y proyectar su éxito y comunicarse constantemente con los usuarios para establecer un flujo de retroalimentación, En esto consiste el Mantenimiento. Finalmente, es importante sentar las bases para el crecimiento y evolución del Data Warehouse en donde el aspecto clave es manejar el crecimiento y evolución de forma iterativa utilizando el Ciclo de Vida propuesto, y establecer las oportunidades de crecimiento y evolución en orden por nivel prioridad.

4.9 Diseño y Arquitectura Tecnica

El área de arquitectura técnica cubre los procesos y herramientas que se aplican a los datos. En el área técnica existen dos conjuntos que tienen distintos requerimientos, brindan sus propios servicios y componentes de almacenaje de datos, por lo que se consideran cada uno aparte: El back room (habitación trasera) y el front room (habitación frontal). El back room es el responsable de la obtención y preparación de los datos, por lo que también se conoce como adquisición de datos y el front room es responsable de entregar los datos a la comunidad de usuario y también se le conoce como acceso de datos.

5. Marco Teorico

5.1 Filtrado Contenido

El filtro de contenido web es una solución de software y/o hardware que tiene como finalidad actuar como un intermediario entre los accesos de los colaboradores a internet, posibilitando la aplicación de políticas definidas por la empresa.

Esto significa que el acceso a Internet deja de ser hecho directamente por el ordenador y pasa a ser realizado a través de la solución de filtrado de contenido web. Para el usuario esta transacción puede ser transparente, o mediante el uso de credenciales para el reconocimiento del usuario durante el primer intento de acceso a Internet. Es importante resaltar que la conexión con el sitio remoto es hecha por el filtro de contenido web, que aplica las reglas, y si el intento de acceso está en conformidad el sitio se brinda al navegador.

Este modelo garantiza que las peticiones sean tunelizadas y, conociendo detalles de quien está realizando la solicitud, puedan ser permitidas o no. Esto garantiza también, en muchos casos, la protección contra phishing y otras contaminaciones a través de páginas en Internet.

Existen diversas soluciones de filtrado de contenido web, desde las más simples que permiten crear reglas de acceso con contenidos que pueden

o no ser accedidos, hasta soluciones más complejas y completas, que poseen bases de datos voluminosas con clasificación de sitios basados en categorías de interés.

¿Cómo aumentar la productividad utilizando el filtro de contenido web?

- Después del concepto de filtrado de contenido web, el incremento de productividad en ambientes corporativos se da a través de la implementación de esta solución, que traerá gran visibilidad sobre los accesos y consumo de Internet.
- Una buena estrategia en la implantación de estas soluciones es permitir todo el tráfico, si la empresa no posee ninguna regla anterior, de esta forma es posible identificar el perfil de consumo y posibles abusos por parte de los colaboradores.
- El más importante, y factor crítico de éxito, es buscar entender con usuarios, líderes de sector, o personas en cargo de confianza, lo que es primordial en Internet para la realización de su trabajo. Esto debe funcionar siempre, de lo contrario su negocio será afectado.
- Las medidas radicales que restringen totalmente el Internet para determinados usuarios deben ser muy bien analizadas, porque además de no ser positivo en términos motivacionales, puede estimular a que el colaborador busque alternativas para burlar el acceso.
- Cada empresa tiene sus necesidades particulares de acceso a Internet, y respetar esto es fundamental para aumentar la productividad y la seguridad en su entorno. Busque evitar radicalismos, entender lo que es primordial para el trabajo, pero también crear accesos en determinados horarios para que se pueda leer noticias, accederse a las redes sociales, estimulando el ocio creativo.
- Esto es totalmente posible y saludable con tecnologías de filtrado de contenido web, pues además de poder controlar horarios, direcciones, usuarios, sitios y etc. Es también visible en la mayoría de las soluciones el consumo de Internet. Con base en ello, y detectando abusos, se pueden tomar medidas de concientización y bloqueos.

5.2 Filtrado por Colaborativo

El filtrado colaborativo es una técnica utilizada por los sistemas de recomendación para solventar los problemas derivados de la sobreinformación que los consumidores sufren en Internet. Esta tendencia crece cada día más, debido a su enorme funcionalidad son más los usuarios que se valen de esta herramienta en sus búsquedas.

Antes del nacimiento de Internet el consumidor no tenía ninguna fuente de información salvo la propia publicidad del producto. El mercado ha pasado de esta escasez de información a la saturación de los mismos. En este contexto, surgen los filtrados colaborativos. Las empresas incorporan estas herramientas en su página y los propios usuarios construyen una inteligencia colectiva mediante un sistema de recomendaciones que son luego estudiados y traducidos mediante algoritmos estadísticos.

Una de las empresas pioneras en incorporar esta herramienta dentro de su web fue la famosa tienda online Amazon.com, que informa a sus usuarios de los productos que podían interesarles partiendo de los que ya había clicado.

Existen diferentes tipos de filtrado a la hora de establecer las recomendaciones, se pueden clasificar en cuatro:

- Filtrados basado en contenido: las recomendaciones se hacen según los contenidos que puedan gustar o interesar.
- Filtrados demográficos: se realizan por las características de los usuarios (edad, sexo, estudios...).
- Filtrados colaborativos: las recomendaciones están basadas en las búsquedas con votos positivos de usuarios similares.
- Filtrados híbridos: mezclan los dos o tres de los filtrados anteriores para una mejor experiencia.

Los filtrados colaborativos sirve para hacer predicciones automáticas sobre los intereses de un usuario mediante la recopilación de preferencias o gustos del mismo consumidor u otros consumidores con intereses comunes.

Los sistemas de filtrado poseen muchas variantes con algoritmos que se utilizan para su elaboración:

- Algoritmos de filtrado colaborativo basados en memoria, o algoritmos de vecinos cercanos (Nearest Neighbour): utilizan los datos de recogidos para calcular la similitud entre los usuarios o elementos comunes. Fue de los primeros en usarse y es sencillo y eficaz. Funcionan buscando usuarios con patrones de evaluación similares con el usuario activo, para el que se está haciendo la selección. También utilizan técnicas estadísticas para encontrar vecinos con un historial de búsqueda parecido al usuario actual. Su principal inconveniente es que necesitan un número mínimo de usuarios para realizar la recomendación.

- Algoritmos de filtrado colaborativo basados en Modelo: se utilizan algoritmos de aprendizaje automático para encontrar patrones. Mejora el rendimiento en cuanto a la predicción porque da un fundamento más intuitivo. Funcionan usando las evaluaciones de los usuarios afines para calcular la elección del usuario activo. Primero elaboran un modelo de las búsquedas del usuario pero este proceso necesita un aprendizaje largo e intensivo.

También existen algoritmos híbridos que combinan ambos modelos pero son complejos y costosos de implementar.
Dificultades que podemos encontrar cuando usamos el filtrado colaborativo:

- Escasez de datos: los sistemas de filtrado colaborativo se basan en conjuntos de datos. Si esta muestra de datos es escasa puede ser muy costosa y poco eficaz. En ocasiones un problema común es empezar de cero, ya que no se pueden recopilar preferencias con precisión y fiabilidad.
- Sinónimos: la diversidad de etiquetas con nombres similares a veces no son reconocidos por los sistema de filtrados cuando en realidad el usuario está buscando el mismo elemento y se pierde información.

Por ejemplo: un usuario que busca ordenadores o computadoras, son sinónimos pero el buscador no los relaciona.

- Haters o black sheep: otra dificultad que afecta a los sistemas de filtrados son las opiniones de los usuarios que no están de acuerdo con nada y todas sus recomendaciones son negativas, empeoran la calidad de las filtraciones.
- Shilling attacks: en los sistemas de recomendación cualquiera puede hacer evaluaciones, pudiendo un usuario votar positivamente sólo a sus productos y servicios y dar negativo a sus competidores, falseando la eficacia de esta herramienta.
- Diversidad: los filtros intentan buscar una diversidad para poder recomendar entre múltiples opciones. En ocasiones este filtro van reduciendo esta variedad dando sólo visibilidad a los productos con mayor popularidad.

El filtrado colaborativo es una gran herramienta para mejorar la visibilidad y la mejor forma de dar a conocer nuevos productos a más clientes ¿Quieres continuar aprendiendo técnicas para tu negocio? No dudes y échale un vistazo a nuestro Postgrado en e-Commerce Omnichannel de IEBS Business School donde aprenderás todo lo que necesitas para dar el mejor impulso a tu empresa.

6. Ventajas y Desventajas

5.2 Ventajas de la Metodologia Inmon.

- Tecnología (Camino Superior). Implica tareas relacionadas con software específico, por ejemplo, Microsoft SQL Analysis Services.
- El almacén de datos realmente sirve como la única fuente de verdad para la empresa, ya que es la única fuente para los almacenes de datos y todos los datos en el almacén de datos están integrados.
- Las anomalías de actualización de datos se evitan debido a una redundancia muy baja. Esto hace que el proceso ETL sea más fácil y menos propenso a fallar.

- Los procesos de negocios se pueden entender fácilmente, ya que el modelo lógico representa las entidades de negocios detalladas.
- Muy flexible: a medida que cambian los requisitos del negocio o cambian los datos de origen, es fácil actualizar el almacén de datos ya que una cosa está en un solo lugar.
- Puede manejar diversas necesidades de informes en toda la empresa.

5.2 Desventajas de la Metodologia Inmon.

- El modelo y la implementación pueden volverse complejos con el tiempo, ya que involucra más tablas y combinaciones.
- Necesita recursos que sean expertos en modelado de datos y de la propia empresa. Este tipo de recursos puede ser difícil de encontrar y, a menudo, son costosos.
- La configuración inicial y la entrega tomarán más tiempo, y la administración debe ser consciente de esto.
- Se necesita más trabajo de ETL ya que los almacenes de datos se crean a partir del almacén de datos.
- Un equipo bastante grande de especialistas debe estar presente para gestionar con éxito el medio ambiente (Breslin, 2004).

5.2 Ventajas de la Metodologia Kimball.

- Rápido de configurar y construir, y la primera fase del proyecto de almacenamiento de datos se entregará rápidamente.
- El esquema en estrella puede ser comprendido fácilmente por los usuarios de negocios y es fácil de usar para informes. La mayoría de las herramientas de BI funcionan bien con el esquema en estrella.
- La huella del entorno de almacenamiento de datos es pequeña, ocupa menos espacio en la base de datos y facilita la administración del sistema.
- El rendimiento del modelo de esquema en estrella es muy bueno. El motor de la base de datos realizará una çombinación estrella.[en] la que se creará un producto cartesiano utilizando todos los valores de dimensión y la tabla de hechos se consultará finalmente para las

filas selectivas. Se sabe que esta es una operación de base de datos muy efectiva.

- Un pequeño equipo de desarrolladores y arquitectos es suficiente para que el almacén de datos funcione de manera efectiva (Breslin, 2004).
- Funciona realmente bien para las métricas por departamento y el seguimiento de KPI, ya que los almacenes de datos están orientados a la información por departamento o por proceso empresarial.
- La obtención de detalles, donde una herramienta de BI recorre varios esquemas en estrella para generar un informe, se puede lograr con éxito utilizando dimensiones conformadas.

5.2 Desventajas de la Metodologia Kimball.

- La esencia de la "fuente de la verdad"se pierde, ya que los datos no se integran completamente antes de atender las necesidades de informes.
- Los datos redundantes pueden causar anomalías en la actualización de los datos a lo largo del tiempo.
- Agregar columnas a la tabla de hechos puede causar problemas de rendimiento. Esto se debe a que las tablas de hecho están diseñadas para ser muy profundas. Si se agregan nuevas columnas, el tamaño de la tabla de hechos se vuelve mucho más grande y no tendrá un buen desempeño. Esto hace que el modelo dimensional sea difícil de cambiar a medida que cambian los requisitos del negocio.
- No se pueden manejar todas las necesidades de informes de la empresa porque el modelo está orientado a los procesos de negocios en lugar de a la empresa en su conjunto.
- La integración de datos heredados en el almacén de datos puede ser un proceso complejo.

ETL: Este término viene de inglés de las siglas Extract-Transform-Load que significan Extraer, Transformar y Cargar. Se refiere a los datos en una empresa

7. Diferencias

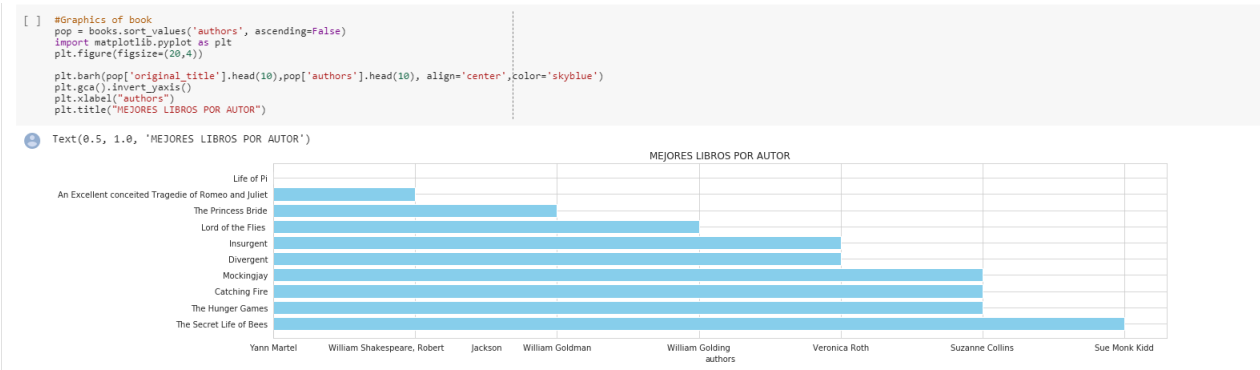
- Según W. H. Inmon (considerado por muchos el padre del Data Warehouse), un Data Warehouse es un conjunto de datos orientados por temas, integrados, variantes en el tiempo y no volátiles, que tienen por objetivo dar soporte a la toma de decisiones.

- Según Ralph Kimball (considerado el principal promotor del enfoque dimensional para el diseño de almacenes de datos), un Data Warehouse es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis.

8. Ejemplos

- Consultas de Mejores Libros por Autor

  Este objetivo se hace necesario para poder iniciar una valoracion de los autores.



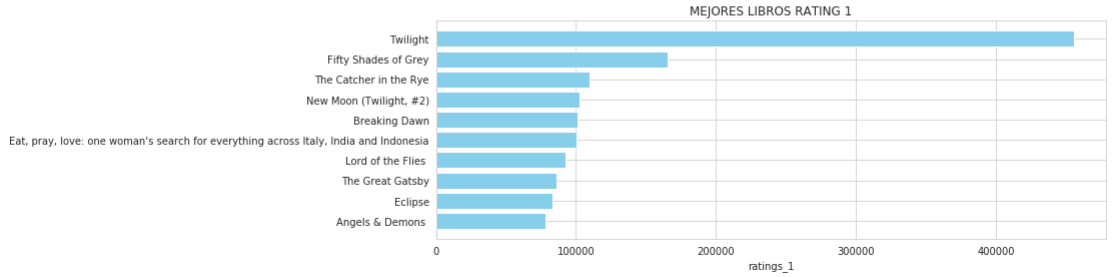
- Consultas de Mejores Libros por Rating 1

  Este objetivo se hace necesario para poder iniciar una valoracion de los libros.

```
[ ]  #Graphics of book
     pop = books.sort_values('ratings_1', ascending=False)
     import matplotlib.pyplot as plt
     plt.figure(figsize=(12,4))

     plt.barh(pop['original_title'].head(10),pop['ratings_1'].head(10), align='center',color='skyblue')
     plt.gca().invert_yaxis()
     plt.xlabel("ratings_1")
     plt.title("MEJORES LIBROS RATING 1")
```

```
Text(0.5, 1.0, 'MEJORES LIBROS RATING 1')
```



9. Conclusion

El método Kimball está orientado a la consulta de la información, por lo que su estructura interna está especialmente diseñada para garantizar una explotación de los datos rápida y sencilla, no requiriendo usuarios especializados para ello. Por el contrario, el método Inmon persigue la integración de todos los datos de la compañía, estando orientado hacia el almacenaje de grandes volúmenes de datos, por lo que su estructura interna normalizada se diseña para evitar la redundancia de datos, simplificar las labores de mantenimiento, etc. cuestiones que complican las consultas de la información, requiriendo que los usuarios finales estén mucho más especializados. Así, podríamos decir que el enfoque de Kimball se ajusta más a proyectos pequeños en los que se persiga un sistema fácilmente explotable y entendible por el usuario y de rápido desarrollo, siendo el modelo de Inmon más apropiado para sistemas complejos de mayor importancia.

Referencias
Referencias

[1]

[2]

[3] http://tdan.com/data-warehouse-design-inmon-versus-kimball/20300

[4] https://blog.bi-geek.com/arquitectura-comparativa-inmon-y-kimball/

[5] https://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimensional/

[6] https://twooctobers.com/blog/8-data-storytelling-concepts-with-examples/

[7] https://www.ucasal.edu.ar/htm/ingenieria/cuadernos/archivos/5-p56-rivadera-formateado.pdf