UNIVERSIDAD PRIVADA DE TACNA
INGENIERIA DE SISTEMAS

TITULO:

Comparativa de las metodologìas de elaboración de Datawarehouse vs metodologias de elaboración de Datalakes

CURSO:

Inteligencia de Negocios

DOCENTE:

Ing. Patrick Cuadros Quiroga

Integrantes:

Maldonado Cancapi, Carlos Alejandro	(2018000660)
Huillca Aroni, Alfredo	(2018060903)
Anahua Huayhua, Jenny Karen	(2018062150)
Coloma Colquehuanca, Kiara	(2018062218)

Tacna - Perú
2022

Comparativa de las metodologìas de elaboración de Datawarehouse vs metodologias de elaboración de Datalakes

May 25, 2022

1. Resumen

En este documento se estudian, analizan y comparan diversas metodologías y herramientas para el desarrollo de un Data Warehouse (DW) y metodologías de elaboración de Datalakes que permita la integración de información en caso, o no que dichos datos se encuentren en diferentes motores de bases de datos y/o provengan de diferentes fuentes de datos, esto, con el fin de convertir los datos en información pertinente y para que dichos datos cumplan con características como la calidad y exactitud, entre otras. Con la gran ventaja de que una vez el Data Warehouse esté desarrollado, se puedan ejecutar procesos de Business Intelligence (BI) para lograr que la información pueda ser usada para la toma de decisiones.

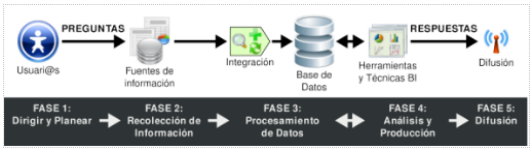
2. Abstract

This document studies, analyzes and compares various methodologies and tools for the development of a Data Warehouse (DW) and methodologies for the elaboration of Datalakes that allow the integration of information in case, or not, that said data is found in different database engines. of data and/or come from different data sources, this, in order to convert the data into relevant information and so that said data meets characteristics such as quality and accuracy, among others. With the great advantage that once the Data Warehouse is developed, Business Intelligence (BI) processes can be executed so that the information can be used for decision making.

3. Introduccion

Los Almacenes de Datos o Data Warehouse, surgieron en la década del 90 del siglo pasado, conocidos como "una colección de datos orientados a un ámbito (empresa, organización), integrada, no volátil y variante en el tiempo, que ayuda al proceso de los sistemas de soporte a la toma de decisiones". El diseño y construcción de los almacenes de datos están ganando cada vez mayor popularidad en las organizaciones, al considerar las ventajas que involucra el análisis de los datos históricos de forma multidimensional para apoyar el proceso de toma de decisiones., resultando complejo en este proceso la recolección de requerimientos, el análisis y el diseño porque no siempre se emplea la metodología adecuada. La definición de metodología para algunos autores es: "una colección de procedimientos, técnicas, herramientas y documentos auxiliares que ayudan a los desarrolladores de software en sus esfuerzos por implementar nuevos sistemas de información". Una metodología está formada por fases, cada una de las cuales se puede dividir en sub-fases, que guiarán a los desarrolladores de sistemas a elegir las técnicas más apropiadas en cada momento del proyecto y también a planificarlo, gestionarlo, controlarlo y evalu-

arlo (3-6).



4. Desarrollo

4.1. DATA WAREHOUSE

W.H. Inmon, considerado el padre de las bodegas de datos en el 92, define los Data Warehouse como: "Un sistema orientado al usuario final, integrado, con variaciones de tiempo y sobre todo una colección de datos como soporte al proceso de toma de decisiones". Por otra parte, Ralph Kimball, considerado como uno de los más importantes precursores y padre del concepto Data Warehouse, lo define como: "una copia de los datos de la transacción estructurados específicamente para preguntar y divulgar" .

En la actualidad, con el fin de lograr un mejor rendimiento las organizaciones tienden a gastar su esfuerzo en la maximización de ingresos y minimizar los gastos, para conseguir dicha meta se requieren elementos que deben reflejarse desde los empleados de nivel más bajo hasta los altos mandos ejecutivos, esto se logra con unos arduos y tediosos procesos de análisis estratégicos para mejorar los procesos empresariales basados en las decisiones de quienes están al mando, es por esto que el lograr obtener informes consolidados o detallados de la información de la empresa, resulta vital en la toma de decisiones y es aquí donde entran en juego los DW, debido a que su propósito en simples cuentas es obtener informes detallados de la información de la empresa con el fin de mejorar la toma de decisiones

4.2. METODOLOGIAS

4.2.1 HEFESTO

Con base en comprender cómo una organización puede crear inteligencia de negocios de sus datos, La metodología de Hefesto se divide en cinco fases y se sintetiza de la siguiente manera:

- Fase 1: Dirigir y Planear.

En esta fase inicial es donde se deberán recolectar los requerimientos de información específicos de los diferentes usuarios, así como entender sus diversas necesidades, para que luego en conjunto con ellos se generen las preguntas que les ayudarán a alcanzar sus objetivos.

- Fase 2: Recolección de Información.

Es aquí en donde se realiza el proceso de extraer desde las diferentes fuentes de información de la empresa, tanto internas como externas, los datos que serán necesarios para encontrar las respuestas a las preguntas planteadas en el paso anterior.

- Fase 3: Procesamiento de Datos.

En esta fase es donde se integran y cargan los datos en crudo en un formato utilizable para el análisis. Esta actividad puede realizarse mediante la creación de una nueva base de datos, agregando datos a una base de datos ya existente o bien consolidando la información.

- Fase 4: Análisis y Producción.

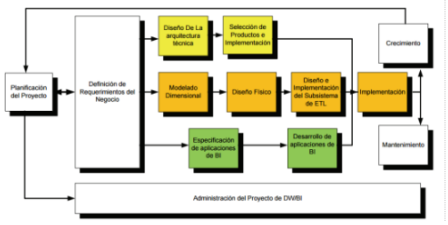
Ahora, se procederá a trabajar sobre los datos extraídos integrados, utilizando herramientas y técnicas propias de la tecnología BI, para crear inteligencia. Como resultado final de esta fase se obtendrán las respuestas a las preguntas, mediante la creación de reportes, indicadores de rendimiento, cuadros de mando, gráficos estadísticos, etc.

- Fase 5: Difusión.

Finalmente, se les entregará a los usuarios que lo requieran las herramientas necesarias, que les permitirán explorar los datos de manera sencilla e intuitiva22

4.2.2 METODOLOGÍA DE RALPH KIMBALL

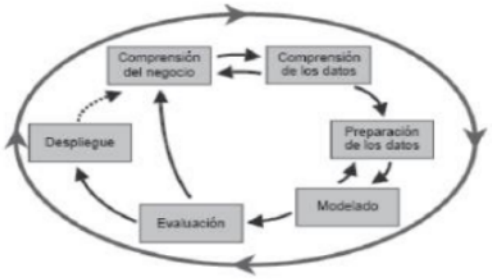
La metodología de Kimball se basa en cuatro principios fundamentales: Centrarse en el negocio: Hay que concentrarse en la identificación de los requerimientos del negocio y su valor asociado, y usar estos esfuerzos para desarrollar relaciones sólidas con el negocio, agudizando el análisis del mismo y la competencia consultiva de los implementadores. Construir una infraestructura de información adecuada: Diseñar 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 elemento 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 avanzado, capacitación, soporte, sitio web y documentación. La construcción de una solución de DW es compleja, y Kimball propone una metodología que ayuda a simplificar dicha solución. Las tareas de esta metodología (ciclo de vida) se muestran en la siguiente figura:



4.2.3 METODOLOGÍA CRISP-DM (Cross-Industry Standard PRocess for Data Mining).

La metodología de CRISP es una de las principales metodologías por seguir por los analistas en la inteligencia de negocios, donde se puede rescatar primordialmente Data Warehouse y Data Mining. La metodología CRISP está sustentada en estándares internacionales que reflejas la robustez de sus procesos y que facilitan la unificación de sus fases en una estructura confiable y amigable para el usuario. Además de ello, esta tecnología interrelaciona las diferentes fases del proceso entre sí, de tal manera que se consolida un proceso iterativo y recíproco. Otro aspecto fundamental de esta tecnología es que es planteada como una metodología imparcial o "neutra respecto a la herramienta que se utilice para el desarrollo del proyecto de Data Warehouse o Data Mining siendo su distribución libre y gratuita.

Ciclo de vida de la metodología CRISP:



El ciclo de vida del proyecto según la metodología según la metodología de CRISP está basado en seis fases cambiantes entre sí y nunca terminantes, lo cual lo postula como un ciclo en constante movimiento.

- Comprensión del negocio: Se trata de entender claramente los requerimientos y objetivos del proyecto siempre desde una visión de negocio. Esta fase se subdivide a su vez en las siguientes categorías: o Definición de los objetivos del negocio (inicial, objetivos de negocio y criterios de éxito del negocio). o Evaluación de la situación (inventario de recursos, requisitos supuestos y requerimientos, riesgos y contingencias, terminología y costes y beneficios). o

Definición de los objetivos del DW (objetivos y criterios de éxito). o Realización del plan del proyecto (plan del proyecto y valoración inicial de herramientas y técnicas).

- Comprensión de los datos: Es conseguir y habituarse con los datos, reconocer las dificultades en la calidad de los datos y reconocer también las fortalezas de estos mismos que pueden servir en el proceso de análisis. Sus subdivisiones son:

- Recolección inicial de datos (informe de recolección).
- Descubrimiento de los datos (informe descriptivo de los datos).
- Exploración de los datos (informe de exploración de los datos).
- Verificación de la calidad de los datos (informes de calidad).

- Preparación de los datos: Es analizar los datos realmente importantes en el proceso de selección, depuración y transformación. Sus subdivisiones son:

- Selección de los datos (motivos para incluirlos o excluirlos).
- Depuración de los datos (reporte de depuración).
- Estructuración de los datos (generación de atributos y registros)
- Integración de los datos (agrupar los datos).
- Formateo de datos (informe de la calidad de datos formateados).

- Modelado: Es la aplicación de técnicas de modelado o de Data Warehouse. Sus subdivisiones son:

- Selección de la técnica de modelado (técnica y sus supuestos).
- Generar el plan de pruebas (plan de pruebas).
- Construcción del modelo (parámetros escogidos, modelos, descripción de los modelos).
- Evaluación del modelo (evaluar el modelo, revisión de los parámetro elegidos).

- Evaluación: Esta fase es muy importante y decisiva, pues corresponde a la evaluación de la escogencia de los modelos anteriores y la toma de decisión respecto a si realmente son útiles en el proceso. Sus subdivisiones son:

- Evaluar resultados (valoración de los resultados respecto al éxito del negocio, modelos aprobados).
- Proceso de revisión (revisar el proceso).
- Determinación de los pasos siguientes (listado de posibles acciones, técnica modelada).

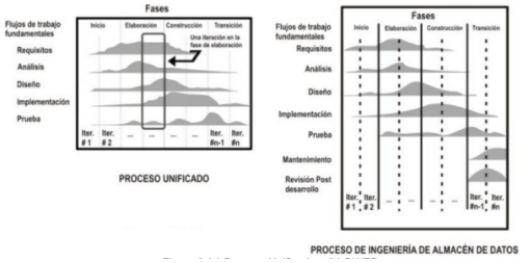
- Despliegue o divulgación: Es la fase de implementación o de divulgación de los modelos anteriormente escogidos y evaluados. Sus subdivisiones son:

- Plan de divulgación o implementación (plan de implementación).
- Plan de monitoreo y mantenimiento (plan de monitoreo y mantenimiento).
- Presentación del informe fina (informe final, presentación final).
- Revisión del proyecto (documentación de la experiencia).

4.2.4 DWEP (Data Warehouse Engineering Process)

Está basada en el proceso unificado estándar aceptado en el ámbito científico e industrial para el desarrollo de software; entre sus principales características se encuentra que es iterativo e incremental, se basa en cuatro fases de desarrollo y siete flujos de trabajo, en la Figura 3 se presentan gráficamente la relación existente entre los flujos de trabajo y las fases tanto del UP como de DWEP, está basado en componentes, utiliza el UML (Unified Modeling Language - Lenguaje Unificado de Modelado) como lenguaje para modelado gráfico es orientada a objetos, independiente de cualquier implementación específica, ya sea relacional o multidimensional y

permite la representación de todas las etapas del diseño de un almacén de datos.



4.2.5 SEMMA

Se define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. Su nombre es el acrónimo correspondiente a las cinco fases básicas del proceso (muestro (sample), explotación (explore), modificación (modify), modelado (model), valoración (assess)) (28).

4.2.6 P3TQ (Product, Place, Price, Time, Quantity)

Está compuesta por dos modelos, el Modelo de Negocio y el Modelo de Explotación de Información. El Modelo de Negocio proporciona una guía de pasos para identificar un problema de negocio o la oportunidad del mismo. El Modelo de Explotación de Información proporciona una guía de pasos para la ejecución de los modelos de explotación de información de acuerdo al modelo identificado en Modelo del Negocio.

4.2.7 KM-IRIS

Fue elaborado por el grupo de Integración y Re-Ingeniería de Sistemas (IRIS) de la Universidad Jaume. Se crea con el objetivo de dirigir el proyecto de desarrollo de un sistema de gestión del conocimiento, consta de cinco fases: identificar, extraer, procesar, almacenar y compartir. Esta metodología pretende cubrir el ciclo completo en el desarrollo de un sistema de gestión del conocimiento. Es una metodología poco difundida y con escasa documentación (16, 27). En la tabla 1 se muestra una breve descripción de sus fases.

4.2.8 Rapid Warehousing Methodology

Es una metodología iterativa que está basada en el desarrollo incremental de un almacén de datos dividido en cinco fases. Esta metodología no incluye lo relativo a técnicas de análisis de la información, por lo que con su aplicación solo se obtendría el almacén de datos y no los multianálisis de los datos para apoyar la toma de decisión.

4.3. DATA LAKE

Es un repositorio de almacenamiento que contiene una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario. Se trata de guardar los datos con el objeto de que puedan ser procesados y utilizados en el momento en que sea necesario. Cada elemento del Data Lake recibe un identificador y etiquetas de metadatos extendidas, con el fin de que 20 pueda ser identificado y recuperado fácilmente. (Fang, 2015) Así, en el Data Lake pueden tener cabida muchos tipos de datos distintos, de diversas fuentes y en diferentes formatos. Esto exige, por supuesto, que la capacidad de almacenamiento sea enorme. En resumen, un sistema Data Lake permite retener todos los datos in procesamiento, dar soporte para todo tipo de perfiles de usuarios, tanto para modelos empresariales como científicos, de esta manera el acceso a la información original es más directa y reduce los pasos necesarios para su procesamiento, con una estructura de datos no definida hasta que los datos no son necesarios.

Enfoque ágil: En un enfoque ágil es requerido para el diseño e implementación de data lake (Data Kitchen, 2020) según data Kitchen, dado que se afrontan las siguientes dificultades:

- Demasiado trabajo
- Errores de datos
- Mala data arruina buenos reportes

- Cambios constantes de necesidades de negocio
- Multitud de herramientas
- Mantenimiento de tuberia de datos nunca termina
- Fatiga de procesamiento manual
- Tareas desalentadoras en migración a nube

En la ejecución de proyectos de analítica de datos, se consideran tanto el ciclo de vida de los datos como el ciclo de vida de valor de negocio, partiendo de implementar una solución, pero afrontar el continuo cambio, como se observa en la siguiente tabla:

Sets de datos	**Valor de negocio**
Organizados	Modelos
Calidad verificada	Gráficas y Tableros de mando
Solicitudes de datos	**Solicitudes de negocio**
Más sets de datos	Preguntas
Reglas de Negocios	Actualizaciones

Lo cual requiere que enfoques ágiles sean tenidos en cuenta, en cuanto a metodología, procesos y herramientas, considerando:

- Capacidad de realizar despliegues hacia ambientes de producción de manera rápida y segura.
- Responder rápidamente a las solicitudes de integración de datos.
- Responder a los cambios de lógica de negocio.
- Tomar acciones correctivas sobre errores en datos oportunamente.
- Automatización de procesamiento y calidad.
- No quedar bloqueado en un enfoque centrado en herramientas o por restricciones de operación de herramientas.

5. Conclusiones

Los lineamientos de diseño son la parte más importante de un proceso para el desarrollo de un DW, con esto bien especificado, el DW difícilmente falle en su funcionamiento y usando el esquema constelación, se tendrá una bonificación a nivel organizativo debido a su cualidad de ser específico.

6. Recomendaciones

Se recomienda tener en cuenta que no todas las herramientas y metodologías no fueron estudiadas y comparadas, existen más de estas para ser analizadas y algunas pueden cumplir mejor para el desarrollo de un DW que las mencionadas en este documento. Una vez se haya desarrollado un DW usando estas comparaciones, ya se podrán aplicar los procesos de BI que normalmente se aplican.

References

[1] Data Lake vs Data Warehouse. Veamos sus principales diferencias. (2022). Powerdata.es. https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/data-lake-vs-data-warehouse.-veamos-sus-principales-diferencias

[2] Gorini, M. (2022). ¿Cuál es la diferencia entre un data lake y un data warehouse? Bismart.com. http://blog.bismart.com/diferencia-entre-data-lake-y-data-warehouse

[3] Gigliotti, M. (2019). Data Lake y Data Warehouse: ¿Qué son y en qué se diferencian? Techedgegroup.com. https://www.techedgegroup.com/es/blog/data-lake-data-warehouse-definicion-diferencias

[4] Data Lake vs Data Warehouse: Key Differences - Talend. (2022). Talend - a Leader in Data Integration Data Integrity. https://www.talend.com/resources/data-lake-vs-data-warehouse/

[5] Emilio Fernández Lastra. (2018, October 10). Data Warehouse y Data Lake. Qué son y para qué sirven. Artyco | the Data Driven Company. https://artyco.com/data-warehouse-data-lake-que-es/

[6] Prakash, S. S. (2020, April). Evolution of Data Warehouses to Data

Lakes for Enterprise Business Intelligence. ResearchGate; unknown. https://www.researchgate.net/publication/343219651EvolutionofDataWarehousestoDataLakesforEnterpriseBusiness

[7] Flores, A. (n.d.). Construyendo y gobernando Data Lakes y Data Warehouses modernos en AWS. Retrieved April 5, 2022, from https://d1.awsstatic.com/events/Summits/AMER2019/Mexico-City/BuildingandgoverningmoderndatalakesanddatawarehousesADB201.pdf

[8] Mendez, A., Britos, A., Garcia-Martínez, P. Y. (2003). Fundamentos de Data Warehouse. Reportes Técnicos En Ingeniería Del Software, 5(1), 19–26. http://artemisa.unicauca.edu.co/ ecaldon/docs/bd/fundamentosdedatawarehouse.pdf

[9] Agudelo, J. (2020) Data Lakes: Aplicaciones, Herramientas y Arquitecturas. Monografía presentada como requisito para optar al Título de Ingeniero de Sistemas y Computación https://repositorio.utp.edu.co/server/api/core/bitstreams/5f56e572-d416-487e-a6d5-ec3a8e45da46/content

[10] Pegdwendé Sawadogo, Jérôme Darmont. On data lake architectures and metadata management. Journal of Intelligent Information Systems, Springer Verlag, 2021, 56 (1), pp.97-120. ff10.1007/s10844- 020-00608-7ff. ffhal-03114365f https://hal.archives-ouvertes.fr/hal-03114365/document