Sistematización y documentación de datos masivos mediante métodos de analítica

Breve descripción:

Este componente aborda las metodologías y prácticas para la sistematización y documentación efectiva de datos masivos en contextos analíticos. Explora desde la creación de documentación técnica hasta la gestión del conocimiento organizacional, incluyendo técnicas de comunicación y mejores prácticas. Proporciona herramientas fundamentales para garantizar la trazabilidad y aprovechamiento del conocimiento en proyectos de datos.

Noviembre 2024

Tabla de contenido

Introducción

En la era de los datos masivos, la sistematización y documentación efectiva del conocimiento se ha convertido en un elemento diferenciador para las organizaciones. La capacidad de capturar, organizar y comunicar el conocimiento técnico determina no solo la eficiencia operativa actual, sino también la capacidad de evolución y mejora continua de los sistemas analíticos.

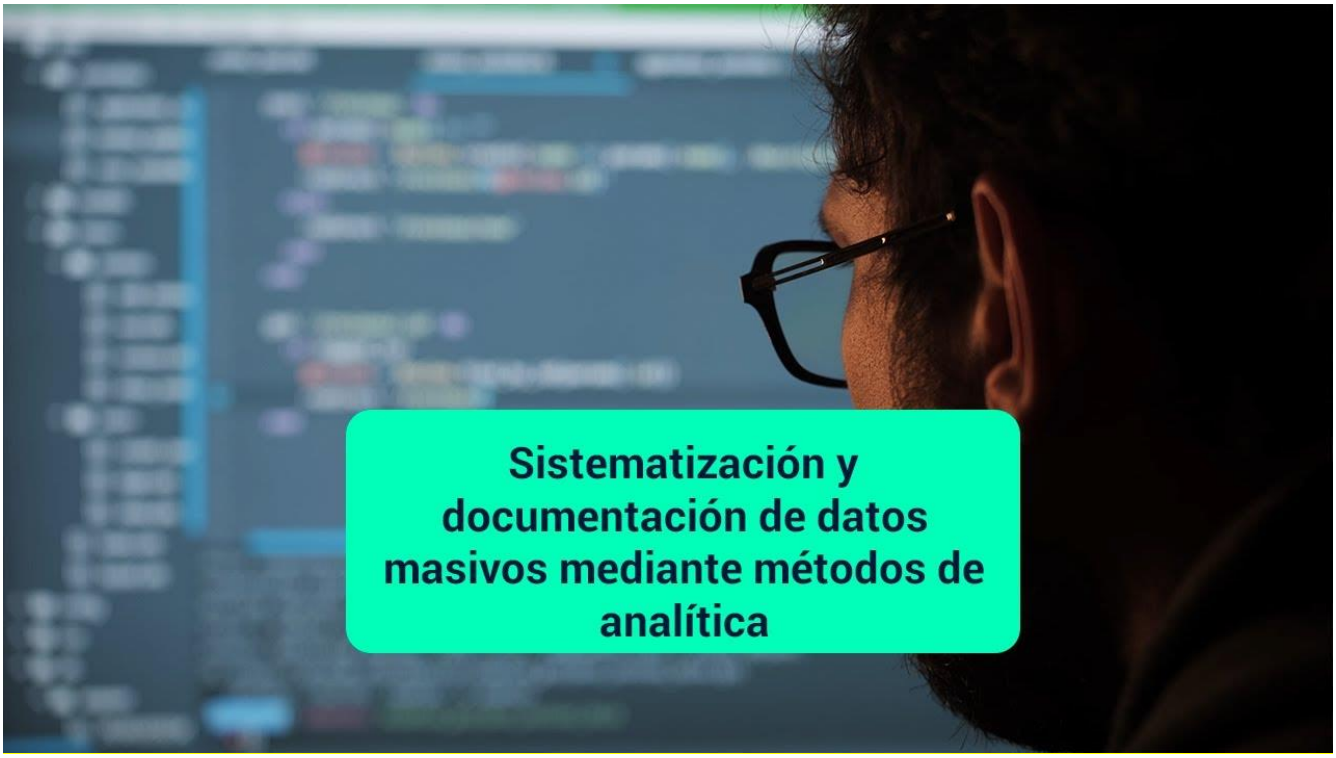
Este componente formativo aborda los aspectos fundamentales de la documentación y sistematización de datos masivos, proporcionando un marco integral para la gestión del conocimiento técnico. Desde la creación de documentación robusta hasta la implementación de sistemas de gestión del conocimiento, cada tema se explora con un enfoque práctico y orientado a resultados.

La documentación técnica moderna trasciende la simple recopilación de información, convirtiéndose en un activo estratégico que facilita la colaboración, el mantenimiento y la evolución de los sistemas de datos. Las metodologías y herramientas presentadas reflejan las mejores prácticas actuales en la industria, adaptadas a las necesidades específicas de proyectos de datos masivos.

Como suele decirse en el campo de la gestión del conocimiento: "La información solo se convierte en conocimiento cuando está documentada, accesible y puede ser aplicada". Este principio guiará nuestra exploración de las técnicas y metodologías presentadas en este componente.

¡Bienvenido al mundo de la sistematización y documentación efectiva de datos masivos!

Video 1. Sistematización y documentación de datos masivos mediante métodos de analítica



Enlace de reproducción del video

Síntesis del video: Sistematización y documentación de datos masivos mediante métodos de analítica
La documentación y sistematización efectiva representan pilares fundamentales en cualquier proyecto de datos masivos. Este componente te guiará a través de las metodologías y mejores prácticas para capturar, organizar y comunicar el conocimiento técnico en entornos analíticos.

Comenzaremos explorando los tipos de documentación técnica y los estándares que garantizan su calidad y utilidad. La documentación bien estructurada es la base para el mantenimiento y evolución de cualquier sistema de datos.

Los informes técnicos avanzados requieren ir más allá de la redacción técnica. Aprenderás metodologías para estructurar y gestionar documentación compleja, asegurando su utilidad para diferentes audiencias.

Las técnicas de comunicación efectiva te permitirán transmitir hallazgos complejos de manera clara y accionable. La narrativa de datos se convierte en una herramienta poderosa para generar impacto con tus análisis.

La gestión del conocimiento cierra el ciclo, proporcionando frameworks para organizar y mantener el conocimiento organizacional a largo plazo. El control de versiones y las mejores prácticas de mantenimiento aseguran la evolución continua de tu base de conocimiento.

Las tendencias actuales apuntan hacia la automatización y la integración continua en la documentación, pero el factor humano sigue siendo esencial para garantizar la calidad y relevancia del conocimiento capturado.

Este componente te proporcionará las herramientas necesarias para implementar sistemas robustos de documentación y gestión del conocimiento en tus proyectos de datos masivos.

¡Bienvenido al arte de la documentación efectiva!

1. Documentación técnica

La documentación técnica representa un elemento fundamental en la gestión de datos masivos y analítica. Más allá de un simple registro de procedimientos, constituye la memoria institucional de los procesos de datos, facilitando la reproducibilidad, el mantenimiento y la transferencia de conocimiento. Este capítulo explora los diferentes tipos de documentación, estándares actuales y herramientas que facilitan esta labor esencial.

1.1. Tipos de documentación

La documentación técnica en el contexto de datos masivos abarca diversos niveles y propósitos. Los documentos de arquitectura proporcionan una visión general de la estructura y flujo de datos, mientras que la documentación de procesos detalla los pasos específicos en las transformaciones y análisis. Los manuales de usuario y las guías de mantenimiento completan el espectro, asegurando que todos los stakeholders puedan interactuar efectivamente con los sistemas de datos.

Para comprender mejor los diferentes tipos de documentación y sus características específicas, consideremos la siguiente clasificación detallada:

Tabla 1. Tipos de documentación y características

Tipo de documentación	Audiencia principal	Contenido clave	Frecuencia de actualización	Nivel de detalle
Arquitectura de datos.	Arquitectos, Ingenieros.	Diagramas de flujo, modelos de datos, dependencias.	Baja - en cambios mayores.	Alto nivel, enfoque en relaciones.

Tipo de documentación	Audiencia principal	Contenido clave	Frecuencia de actualización	Nivel de detalle
Procesos ETL.	Desarrolladores, Analistas.	Transformaciones, reglas de negocio, validaciones.	Media - con cada modificación.	Detallado, paso a paso.
APIs e interfaces.	Desarrolladores externos e internos.	Endpoints, parámetros, ejemplos de uso.	Alta - con cada cambio de API.	Técnico y específico.
Calidad de datos.	Analistas, Gestores de datos.	Métricas, umbrales, procedimientos de validación.	Alta - seguimiento continuo.	Medio, énfasis en métricas.
Seguridad y acceso.	Administradores, Auditores.	Políticas, permisos, protocolos.	Media - cambios en políticas.	Exhaustivo en procedimientos.
Guías de usuario.	Usuarios finales.	Instrucciones de uso, casos comunes, solución de problemas.	Media - nuevas funcionalidades.	Básico, enfoque práctico.

Fuente. OIT, 2024.

Esta clasificación permite alinear cada tipo de documentación con sus objetivos específicos y audiencia prevista, facilitando la creación y mantenimiento de una estructura documental coherente y efectiva.

1.2. Estándares y mejores prácticas

Los estándares de documentación técnica han evolucionado significativamente con la expansión de los sistemas de datos masivos. La adopción de formatos estandarizados como Markdown o AsciiDoc facilita la creación y mantenimiento de documentación que puede versionarse junto con el código. La documentación como

código (Documentation as Code) ha emergido como un paradigma que permite tratar los documentos técnicos con el mismo rigor y herramientas que el código fuente.

Las mejores prácticas actuales enfatizan la importancia de la documentación viva, que evoluciona junto con los sistemas que describe. Esto implica integrar la documentación en el ciclo de desarrollo, estableciendo procesos de revisión y actualización regulares. La automatización juega un papel importante, permitiendo generar parte de la documentación directamente desde el código y los metadatos del sistema.

La consistencia en el estilo y formato resulta esencial para la usabilidad de la documentación. El uso de plantillas predefinidas y guías de estilo ayuda a mantener la coherencia a través de diferentes equipos y proyectos. La documentación debe ser concisa pero completa, evitando tanto la sobrecarga de información como las omisiones significativas.

1.3. Herramientas de documentación

El ecosistema de herramientas para documentación técnica se ha expandido considerablemente en los últimos años. Los sistemas modernos de documentación combinan capacidades de edición colaborativa, control de versiones y publicación automatizada. Herramientas como Sphinx, MkDocs y Confluence han establecido nuevos estándares en la creación y mantenimiento de documentación técnica.

La integración con sistemas de control de versiones como Git permite mantener la documentación sincronizada con el código y los procesos que describe. Los sistemas de integración continua pueden automatizar la validación y publicación de

documentación, asegurando que los cambios en el código se reflejen adecuadamente en la documentación correspondiente.

Las herramientas de documentación API, como Swagger o OpenAPI, facilitan la creación y mantenimiento de documentación específica para interfaces de programación. Estas herramientas pueden generar automáticamente documentación actualizada a partir de las definiciones de API, reduciendo el riesgo de discrepancias entre la implementación y la documentación.

La selección de herramientas debe considerar factores como la escalabilidad, la facilidad de uso para los contribuyentes, las capacidades de búsqueda y la integración con flujos de trabajo existentes. La adopción de herramientas que soportan formatos estándar y exportación a múltiples formatos aumenta la flexibilidad y longevidad de la documentación.

2. Informes técnicos avanzados

Los informes técnicos avanzados constituyen una herramienta fundamental para comunicar resultados, metodologías y hallazgos en proyectos de analítica de datos. La capacidad de transmitir información técnica compleja de manera clara y estructurada determina en gran medida el impacto y la utilidad de los análisis realizados. Este capítulo explora las metodologías y mejores prácticas para la creación y gestión de informes técnicos efectivos.

2.1. Estructura y organización

La estructuración efectiva de un informe técnico requiere un balance entre rigor metodológico y claridad expositiva. La organización jerárquica de la información permite a los lectores navegar el contenido según sus necesidades específicas, desde resúmenes ejecutivos hasta detalles técnicos profundos.

Para comprender mejor los elementos constitutivos de un informe técnico avanzado y su propósito, consideremos la siguiente estructura estándar y sus componentes:

Tabla 2. Elementos que integran un informe técnico avanzado

Sección	Contenido principal	Audiencia objetivo	Nivel de detalle técnico	Extensión recomendada
Resumen ejecutivo.	Hallazgos clave y recomendaciones.	Gerencia y tomadores de decisiones.	Bajo - enfoque en impacto.	1-2 páginas.
Introducción.	Contexto, objetivos y alcance.	Todos los stakeholders.	Medio - marco general.	2-3 páginas.

Sección	Contenido principal	Audiencia objetivo	Nivel de detalle técnico	Extensión recomendada
Metodología.	Técnicas y herramientas utilizadas.	Equipo técnico y analistas.	Alto - detalles específicos.	3-5 páginas.
Análisis de datos.	Procesamiento y resultados.	Analistas y expertos del dominio.	Alto - análisis detallado.	5-10 páginas.
Visualizaciones.	Gráficos y dashboards.	Todos los stakeholders.	Medio - interpretación clara.	3-5 páginas.
Conclusiones.	Interpretaciones y recomendaciones.	Gerencia y equipo técnico.	Medio - síntesis técnica.	2-3 páginas.
Anexos técnicos.	Código, logs y documentación adicional.	Equipo técnico.	Muy alto - detalles completos.	Sin límite.

Fuente. OIT, 2024.

Esta estructura proporciona un marco de referencia que puede adaptarse según las necesidades específicas del proyecto y la organización, manteniendo siempre un flujo lógico y coherente de información.

2.2. Metodologías de documentación

Las metodologías de documentación para informes técnicos avanzados han evolucionado para adaptarse a las necesidades de proyectos de datos cada vez más complejos. El enfoque modular permite desarrollar secciones independientes que luego se integran en un documento cohesivo, facilitando la colaboración entre diferentes equipos y especialistas.

La creación de informes técnicos sigue un proceso iterativo que involucra múltiples etapas y participantes. La siguiente infografía ilustra el flujo de trabajo típico en la generación de informes técnicos avanzados, desde la recopilación inicial de información hasta la publicación final.

Figura 1. Ciclo de vida del informe técnico

Fuente. OIT, 2024.

El diagrama representa las etapas interconectadas del proceso de documentación, destacando los puntos de revisión, validación y retroalimentación. Esta

visualización ayuda a comprender cómo cada fase contribuye a la calidad final del informe, donde los diferentes roles interactúan durante el proceso.

2.3. Gestión de versiones

La gestión de versiones en informes técnicos trasciende el simple control de cambios. Implica mantener un registro detallado de la evolución del documento, las decisiones tomadas y las modificaciones realizadas. Esta práctica resulta especialmente importante en entornos colaborativos donde múltiples autores contribuyen al documento.

Los sistemas modernos de control de versiones permiten mantener un historial completo de cambios, facilitando la trazabilidad y la capacidad de revertir modificaciones cuando sea necesario. La integración con sistemas de gestión de código permite vincular versiones específicas del informe con los conjuntos de datos y análisis correspondientes.

El versionado semántico (Semantic Versioning) ha emergido como un estándar efectivo para la numeración de versiones de informes técnicos. Este sistema utiliza tres números (MAYOR. MENOR. PARCHE) para indicar la naturaleza de los cambios realizados, facilitando la comprensión rápida del impacto de cada actualización.

La documentación de cambios entre versiones debe incluir no solo qué se modificó, sino también por qué se realizaron los cambios. Esta información contextual resulta invaluable para comprender la evolución del análisis y las decisiones tomadas en diferentes momentos del proyecto.

3. Técnicas de comunicación

La comunicación efectiva de resultados analíticos representa uno de los mayores retos en proyectos de datos masivos. La capacidad de traducir análisis complejos en narrativas comprensibles determina el valor práctico de todo el trabajo analítico. Este capítulo aborda las técnicas y estrategias para comunicar resultados técnicos de manera efectiva a diferentes audiencias.

3.1. Presentación de resultados

La presentación de resultados analíticos requiere un equilibrio delicado entre precisión técnica y claridad expositiva. El primer paso consiste en identificar la audiencia objetivo y adaptar el nivel de detalle técnico según sus necesidades y conocimientos previos.

Los resultados deben presentarse de manera estratificada, comenzando con los hallazgos más significativos y progresando hacia detalles más específicos. Esta estructura permite a cada miembro de la audiencia profundizar hasta el nivel que resulte relevante para sus necesidades.

Algunos elementos clave para una presentación efectiva incluyen:

- Resumen ejecutivo que destaque los hallazgos principales y su impacto.
- Contextualización clara del problema y los objetivos del análisis.
- Metodología explicada en términos comprensibles.
- Ejemplos concretos que ilustren los conceptos abstractos.

3.2. Comunicación visual

La comunicación visual transforma datos abstractos en representaciones intuitivas que facilitan la comprensión de patrones y relaciones complejas. La selección del tipo de visualización debe basarse en la naturaleza de los datos y el mensaje que se desea transmitir.

Los principios de diseño visual juegan un papel fundamental en la efectividad de la comunicación. El uso apropiado del color, la disposición espacial y la jerarquía visual guían la atención del espectador hacia los elementos más relevantes. La simplicidad y la claridad deben priorizarse sobre la complejidad visual.

En el contexto de datos masivos, las visualizaciones interactivas permiten a los usuarios explorar los datos según sus intereses específicos. Sin embargo, cada elemento interactivo debe tener un propósito claro y añadir valor a la comprensión del análisis.

3.3. Narrativa de datos

La narrativa de datos transforma números y estadísticas en historias coherentes que resuenan con la audiencia. Una narrativa efectiva conecta los datos con su contexto empresarial o social, haciendo explícito su impacto y relevancia.

La estructura narrativa debe seguir un arco lógico que guíe al lector desde el planteamiento del problema hasta las conclusiones y recomendaciones. Cada elemento de la historia debe construirse sobre los anteriores, creando una progresión natural que facilite la comprensión.

Los elementos fundamentales de una narrativa de datos incluyen:

- Un contexto claro que establezca la importancia del análisis.

- Una progresión lógica de hallazgos que construya el argumento.
- Puntos de inflexión que destaquen descubrimientos significativos.
- Conclusiones que conecten los hallazgos con acciones concretas.

La narrativa debe mantener un equilibrio entre el rigor analítico y la accesibilidad. El uso de analogías y ejemplos del mundo real puede ayudar a explicar conceptos técnicos complejos sin sacrificar la precisión.

La documentación que acompaña la narrativa debe proporcionar detalles suficientes para que otros analistas puedan reproducir y validar los resultados. Esto incluye referencias a las fuentes de datos, descripción de las transformaciones realizadas y código utilizado en el análisis.

El éxito de una narrativa de datos se mide por su capacidad para impulsar acciones basadas en los hallazgos presentados. Por ello, las recomendaciones deben ser específicas, accionables y estar respaldadas por la evidencia presentada en el análisis.

La retroalimentación de la audiencia constituye un elemento valioso para refinar y mejorar la narrativa. La capacidad de adaptar la presentación según las preguntas y comentarios recibidos demuestra dominio del tema y fortalece la credibilidad del análisis.

4. Gestión del conocimiento

La gestión del conocimiento en el contexto de datos masivos representa un desafío que va más allá del simple almacenamiento de información. Implica la creación de un ecosistema que facilite la captura, organización, distribución y evolución del conocimiento organizacional relacionado con datos y análisis. Este capítulo explora las estrategias y mejores prácticas para gestionar efectivamente el conocimiento en proyectos de analítica avanzada.

4.1. Organización de la información

La organización efectiva de la información constituye la base de una gestión del conocimiento exitosa. Un sistema bien estructurado permite no solo almacenar información, sino también facilitar su descubrimiento y utilización cuando se necesita. La arquitectura de información debe considerar tanto las necesidades actuales como la escalabilidad futura.

Para comprender mejor las diferentes dimensiones de la organización de información en proyectos de datos masivos, consideremos la siguiente taxonomía de activos de conocimiento:

Tabla 3. Taxonomía sugerida sobre activos de conocimiento

Tipo de activo	Contenido principal	Método de organización	Frecuencia de actualización	Herramientas recomendadas
Documentación técnica.	Especificaciones, arquitectura, código.	Sistema jerárquico por componentes.	Alta - con cada cambio.	Git, Confluence, Sphinx.
Conocimiento procedimental.	Flujos de trabajo, procedimientos, guías.	Organización por procesos.	Media - cambios en procesos.	Wiki, Notion, Sharepoint.

Tipo de activo	Contenido principal	Método de organización	Frecuencia de actualización	Herramientas recomendadas
Metadatos.	Diccionarios de datos, linaje.	Estructura relacional.	Alta - cambios en datos.	Data catalogs, neo4j.
Mejores prácticas.	Guías, estándares, lecciones aprendidas.	Categorización temática.	Baja - revisión periódica.	Confluence, gitbook.
Artefactos de modelo.	Modelos, pipelines, configuraciones.	Versionamiento semántico.	Alta - iteración continua.	MLflow, DVC, Git LFS.
Conocimiento contextual.	Decisiones, razonamiento, contexto.	Organización temporal y por proyecto.	Media - eventos significativos.	Confluence, Jira.

Fuente. OIT, 2024.

Esta taxonomía proporciona un marco de referencia para organizar sistemáticamente los diferentes tipos de conocimiento generados en proyectos de analítica, facilitando su gestión y acceso efectivo.

4.2. Control de versiones

El control de versiones en la gestión del conocimiento trasciende la mera gestión de cambios en documentos. Implica mantener un registro coherente de la evolución del conocimiento organizacional, incluyendo el contexto y razonamiento detrás de cada modificación.

El proceso de control de versiones en la gestión del conocimiento sigue un flujo complejo que integra múltiples aspectos y consideraciones. La siguiente infografía

ilustra la interrelación entre los diferentes elementos del sistema de control de versiones y su impacto en la gestión del conocimiento organizacional:

Figura 2. Ecosistema de control de versiones en gestión del conocimiento

Fuente. OIT, 2024.

Como se observa en el diagrama, el control de versiones actúa como un sistema integrado que conecta diferentes aspectos de la gestión del conocimiento, desde la captura inicial hasta la distribución y actualización del conocimiento organizacional. Elementos clave en el control de versiones incluyen:

- Políticas de versionamiento claramente definidas.

- Sistemas de nomenclatura estandarizados.
- Procesos de revisión y aprobación.
- Mecanismos de trazabilidad.

4.3. Mejores prácticas de mantenimiento

El mantenimiento efectivo del conocimiento organizacional requiere un enfoque sistemático y proactivo. Las mejores prácticas de mantenimiento aseguran que el conocimiento permanezca relevante, accesible y útil a lo largo del tiempo. La revisión periódica del conocimiento almacenado resulta esencial para mantener su validez y utilidad. Este proceso debe incluir la identificación de contenido obsoleto, la actualización de información desactualizada y la incorporación de nuevo conocimiento relevante.

El mantenimiento del conocimiento debe considerar múltiples dimensiones:

- **Dimensión tecnológica**: asegurar que la infraestructura y herramientas de gestión del conocimiento funcionen eficientemente y se mantengan actualizadas.
- **Dimensión organizacional**: establecer roles y responsabilidades claras para el mantenimiento del conocimiento, incluyendo la designación de expertos en dominios específicos.
- **Dimensión cultural**: fomentar una cultura de compartir y mantener el conocimiento, donde la documentación y actualización de información se consideren parte integral del trabajo.

La automatización juega un papel cada vez más importante en el mantenimiento del conocimiento. Las herramientas modernas pueden ayudar a identificar contenido

desactualizado, verificar enlaces rotos y sugerir actualizaciones basadas en cambios en sistemas relacionados.

La gestión de dependencias entre diferentes elementos de conocimiento requiere especial atención. Cuando se actualiza un componente, es necesario evaluar y actualizar todos los elementos relacionados para mantener la coherencia del sistema de conocimiento.

El monitoreo continuo de la utilización y efectividad del conocimiento proporciona insights valiosos para el mantenimiento. Las métricas de uso, feedback de usuarios y patrones de acceso pueden ayudar a identificar áreas que requieren atención o mejora.

La documentación del proceso de mantenimiento en sí mismo resulta fundamental. Esto incluye:

- Registro de decisiones de mantenimiento.
- Documentación de procedimientos de actualización.
- Historiales de cambios y sus motivaciones.
- Lecciones aprendidas durante el proceso.

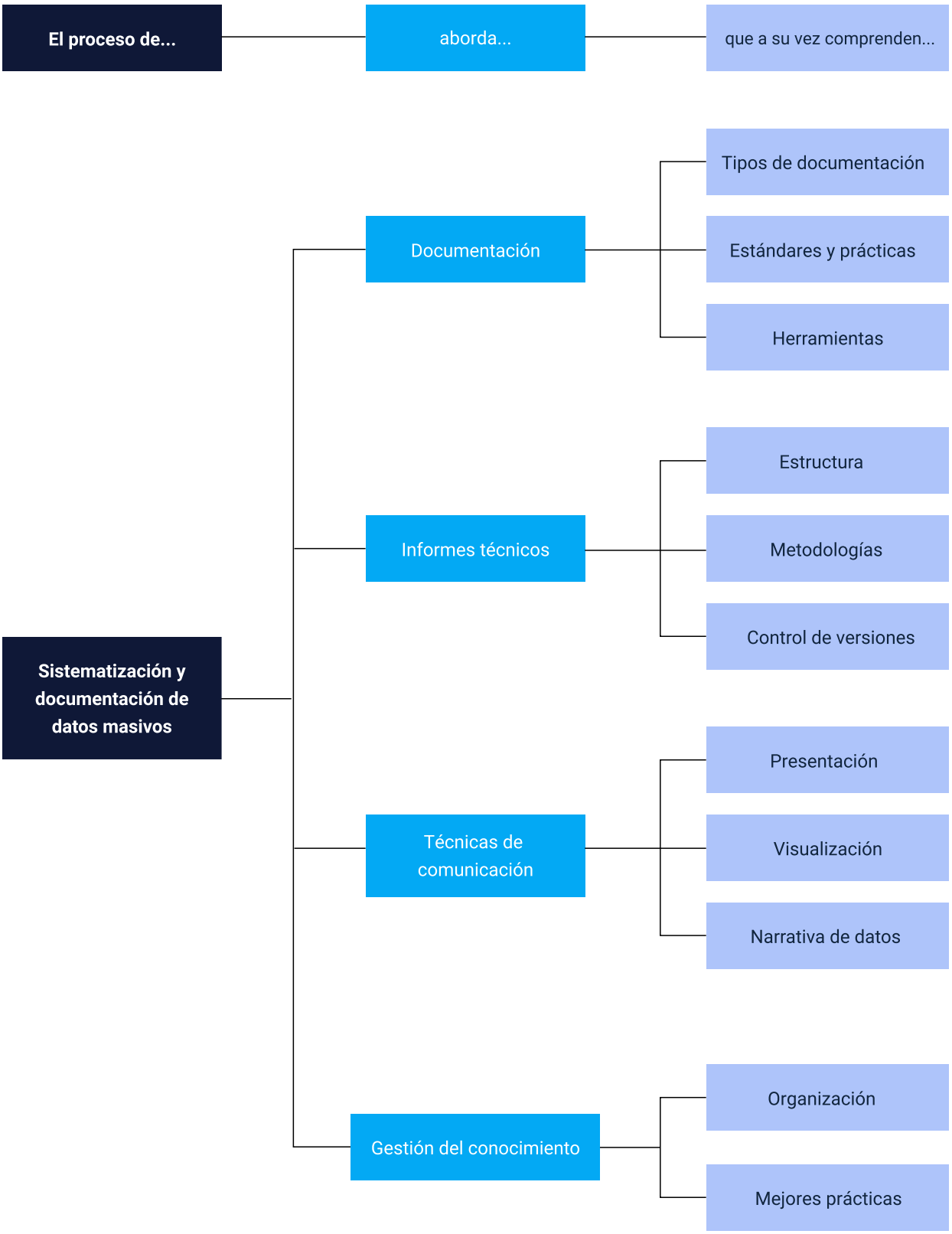
La sostenibilidad a largo plazo del sistema de gestión del conocimiento depende de la capacidad para adaptarse a cambios en las necesidades organizacionales, tecnologías emergentes y prácticas evolutivas en el campo de la analítica de datos.

Síntesis

El diagrama representa la estructura integral del componente formativo sobre sistematización y documentación de datos masivos mediante métodos de analítica. Partiendo del concepto central de sistematización y documentación, se ramifica en cuatro áreas esenciales: documentación, informes técnicos avanzados, técnicas de comunicación y gestión del conocimiento. Cada área incorpora subtemas específicos que constituyen los elementos fundamentales para una gestión efectiva del conocimiento en proyectos de análisis de datos.

Esta organización ilustra el flujo natural del proceso de documentación y gestión del conocimiento, desde la creación inicial de documentación técnica hasta la implementación de sistemas integrales de gestión del conocimiento. La interrelación entre las diferentes áreas muestra cómo cada aspecto se complementa con los demás, creando un ecosistema cohesivo que asegura la captura, organización y transmisión efectiva del conocimiento técnico.

El diagrama funciona como una hoja de ruta visual para comprender la estructura y alcance del componente, permitiendo al aprendiz visualizar rápidamente la progresión del aprendizaje y las conexiones entre los diferentes aspectos de la documentación técnica. Se sugiere utilizarlo como referencia para organizar el estudio y comprender la integración de los diversos elementos que componen un sistema efectivo de documentación y gestión del conocimiento en proyectos de datos masivos.



Fuente. OIT, 2024.

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
1. Modelamiento avanzado de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023e, marzo 27). Modelos y metodologías de analítica.	Video	https://www.youtube.com/watch?v=96pohadjEWE
2. Inteligencia de negocios	Ecosistema de Recursos Educativos Digitales SENA. (2023d, marzo 27). Bodegas de datos	Video	https://www.youtube.com/watch?v=SsP1tA6hAdg
2. Inteligencia de negocios	Ecosistema de Recursos Educativos Digitales SENA. (2023a, marzo 23). Modelos y esquemas de bodega de datos.	Video	https://www.youtube.com/watch?v=Uq6WxfzaroM
3. Análisis exploratorio de datos	Limpiar datos de Excel, CSV, PDF y Hojas de cálculo de Google con el intérprete de datos. (s. f.). Tableau.	Portal web	https://help.tableau.com/current/pro/desktop/es-es/data_interpreter.htm
3. Análisis exploratorio de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023c, marzo 24). Introducción a la aplicación de herramientas estadísticas en la presentación de datos.	Video	https://www.youtube.com/watch?v=M9q9zxX8Evc%3C
4. Preparación avanzada de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023c, julio 25). Procesamiento y análisis de datos.	Video	https://www.youtube.com/watch?v=8OSIN2kdU5o
4. Preparación avanzada de datos	Ecosistema de Recursos Educativos Digitales SENA. (2023e, diciembre 30). Modelamiento, análisis y preparación de datos.	Video	https://www.youtube.com/watch?v=HjJpqHD6sV0

Glosario

Arquitectura estrella: modelo de diseño de bases de datos dimensionales donde una tabla de hechos central se conecta con múltiples tablas de dimensiones desnormalizadas.

Bodega de datos: sistema de almacenamiento diseñado específicamente para el análisis y reporteo, que integra datos de múltiples fuentes en un modelo unificado.

Copo de nieve: variante de la arquitectura estrella donde las dimensiones están normalizadas, creando una estructura más compleja pero con mejor eficiencia de almacenamiento.

Data mart: subconjunto de una bodega de datos enfocado en un área específica del negocio o departamento.

Dimensiones conformadas: tablas de dimensiones estandarizadas que se comparten entre diferentes data marts, asegurando consistencia en el análisis.

ETL avanzado: procesos sofisticados de Extracción, Transformación y Carga que incluyen validaciones complejas y transformaciones avanzadas de datos.

Feature importance: medida que indica la relevancia o contribución de cada variable en un modelo predictivo o análisis estadístico.

Metadatos empresariales: información que describe el contenido, formato, estructura y uso de los datos en un contexto empresarial.

Metodología Inmon: enfoque "top-down" para el diseño de bodegas de datos, que comienza con una visión empresarial completa y luego deriva en data marts específicos.

Metodología Kimball: enfoque "bottom-up" para el diseño de bodegas de datos, que construye data marts incrementalmente que luego se integran en una solución empresarial.

Normalización avanzada: proceso de diseño de bases de datos que va más allá de la tercera forma normal, incluyendo BCNF y formas normales superiores.

Prueba de hipótesis: método estadístico para tomar decisiones sobre poblaciones basándose en muestras de datos.

Reglas de negocio: políticas, condiciones y restricciones que definen cómo se deben gestionar y validar los datos en un contexto empresarial.

Tabla de hechos: tabla central en un modelo dimensional que contiene las métricas o medidas del negocio y las claves foráneas a las dimensiones.

Tablas de dimensiones: tablas que contienen los atributos descriptivos utilizados para analizar los datos en las tablas de hechos.

Transformación de datos: proceso de convertir datos de un formato o estructura a otro, incluyendo limpieza, normalización y agregación.

Validación cruzada: técnica estadística para evaluar modelos analíticos dividiendo los datos en conjuntos de entrenamiento y prueba.

Variables categóricas: tipos de datos que representan categorías o grupos discretos, que pueden ser nominales u ordinales.

Visualización avanzada: técnicas sofisticadas para representar datos complejos de manera visual, incluyendo gráficos interactivos y multidimensionales.

Workflow ETL: flujo de trabajo que define la secuencia y dependencias de los procesos de extracción, transformación y carga de datos.

Referencias bibliográficas

Aguilar, L. J. (2020). Inteligencia de negocios y analítica de datos. Marcombo.

De Pablos Heredero, C., Agius, J. J. L. H., Romero, S. M., & Salgado, S. M. (2019). Organización y transformación de los sistemas de información en la empresa. ESIC.

Díaz, C. O., Soler, P., Pérez, M. & Mier, A. (2024). OMASHU: La ciencia detrás del éxito; Big Data e IA en los eSports. Revista SISTEMAS, 170, 61-79.

Guardelli, E. (2024). Minería de Procesos: Convertir Datos en Valor. MedTechBiz.

Jones, H. (2018). Analítica de Datos: Una guía esencial para principiantes en minería de datos, recolección de datos, análisis de Big Data para negocios y conceptos de inteligencia empresarial. Independently Published.

Maldonado, L. (2012). Data Analysis Using Regression and Multilevel/Hierarchical Models. Persona y Sociedad, 26(1), 191. https://doi.org/10.53689/pys.v26i1.12

McKinsey, W. (2023). Python para análisis de datos. Anaya Multimedia.

Orlandi, M. A. M. (2024). Tecnologías Big Data, Minería de Datos y Analítica aplicada a la gestión de Recursos Humanos: contiene: un caso de estudio. Editora Dialética.

Peraza, E. A. C. (2012). Estructuras y Fundamentos de Datos. Guía de ejercicios prácticos. Lulu.com.

Shovic, J. C. & Simpson, A. (2019). Python All-in-One For Dummies. John Wiley & Sons.

Subirats Maté, L., Pérez Trenard, D. O., Calvo González, M. & Isabel Guitart Hormigo. (2019). Introducción a la limpieza y análisis de los datos. https://openaccess.uoc.edu/bitstream/10609/148647/1/IntroduccionALaLimpiezaYAnalisisDeLosDatos.pdf

Wilke, C. O. (2019). Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures. O'Reilly Media.

Créditos

Elaborado por:

Organización Internacional del Trabajo