Sala Especializada para la Modernización Tecnológica de la Producción Estadística del Consejo Asesor Técnico del Sistema Estadístico Nacional (CASEN) 2023-2025 y Comité de Administración de Datos (CAD)

Seminario "*Arquitectura de datos: fundamentos y buenas prácticas*"
Nota conceptual

Tipo de evento: mixto (presencial y virtual).
Lugar: Auditorio DANE Carrera 59 No. 26-70 Interior I – CAN.
Fecha: 20 de junio de 2024.
Hora: 8:00 a. m. a 12:30 m.
Organizan: Expertos de la Sala Especializada para la Modernización Tecnológica de la Producción Estadística del CASEN y Secretarías Técnicas del CASEN y el CAD.
Enlace: YouTube del DANE.

Objetivo

Compartir con los actores del ecosistema de datos del Sistema Estadístico Nacional (SEN) y con los administradores de datos sectoriales del Estado, elementos de la arquitectura de datos, su importancia y las mejores prácticas para su implementación efectiva en el ciclo de vida de los proyectos de la ciencia de datos, de los proyectos de aprovechamiento de datos y de su articulación para la gestión pública.

Contexto

Hoy los datos son un activo invaluable que generan valor social, económico y ambiental para la política pública y privada, y su adecuada gestión y aprovechamiento son esenciales para impulsar la toma de decisiones basada en evidencia y el desarrollo de soluciones innovadoras. En este contexto, la arquitectura de datos emerge como un pilar fundamental para garantizar una gestión eficiente y efectiva de los datos a lo largo de su ciclo de vida.

Bajo este horizonte, Colombia viene consolidando un marco de política para aumentar el uso y el aprovechamiento de datos, enfocado en:

- La Política Nacional de Explotación de Datos y Big Data con el CONPES 3920 de 2018.
- La Política Nacional de Transformación Digital e Inteligencia Artificial con el CONPES 3975 de 2019.
- La creación del Comité de Administración de Datos (CAD) como instancia del SEN que articula la producción de información estadística con los requerimientos de información para la política pública, resoluciones 0451 de 2020 y la Resolución 2259 de 2023, esta última reglamenta lo dispuesto en la Ley de Estadísticas Oficiales (Ley 2335 de 2023)
- El Plan Nacional de Infraestructura de Datos (PNID) a partir de la política para la reactivación, la repotenciación y el crecimiento sostenible e incluyente -CONPES 4023 de 2020 y la resolución 460 de 2020.
- El Modelo de la Gobernanza de la Infraestructura de Datos y la creación del Comité Nacional de Datos (CND) con el Decreto 1389 de 2022
- La implementación de una estrategia de datos sectorial en el marco de las recomendaciones, las discusiones y los lineamientos expedidos por el Comité Nacional de Datos y el Comité de Administración de Datos con el Plan Nacional de Desarrollo (PND) 2022-2026: "Colombia, potencia mundial de la vida".
- El Marco de Referencia de Arquitectura Empresarial en su tercera versión (MRAE) con la resolución 1978 de 2023.
- El Plan Estadístico Nacional 2023-2027 con la Resolución 2337 de 2023.
- El proyecto de Ley de Datos que cursa su discusión en el Congreso.

El DANE, como entidad rectora del SEN en Colombia reconoce la importancia de fortalecer las capacidades de los administradores de datos sectoriales y de las entidades del SEN en general, en temas de arquitectura de datos. Esta iniciativa responde a la implementación de la hoja de

ruta del PNID y a la necesidad de mantenerse a la vanguardia en la gestión de datos, de aprovechar al máximo el potencial de los proyectos de ciencia de datos en beneficio del país.

En este contexto, el CASEN, a través de su Sala para la Modernización Tecnológica junto con el CAD organizan el Seminario "*Arquitectura de datos: fundamentos y buenas prácticas*" que fortalece las competencias de los participantes en áreas clave como el diseño de modelos de datos, la gobernanza y la calidad de datos, la seguridad y la privacidad, el ciclo de vida de la arquitectura de datos, y las herramientas y las tecnologías asociadas. Además, resaltará la importancia de promover una cultura organizacional que respalde y facilite la adopción de prácticas sólidas de arquitectura de datos.

Al capacitar a los profesionales, a los líderes en estas temáticas, a los administradores de datos y a las entidades del SEN en general, se podrán desarrollar proyectos de aprovechamiento y gestión de los datos y de la ciencia de datos que beneficiarán la toma de decisiones informadas, la optimización de procesos y el desarrollo de soluciones innovadoras que beneficien a la sociedad colombiana.

Síntesis temática

Introducción a la arquitectura de datos

La arquitectura de datos es un componente fundamental en proyectos de ciencia de datos, ya que define cómo se organizan, procesan y utilizan los datos dentro de una organización. Es un conjunto de modelos, políticas, reglas y definiciones que describen la captura, el almacenamiento, la gestión, la integración y el uso de los datos corporativos. Su principal objetivo es establecer los estándares y los lineamientos para gobernar los datos como un activo valioso y asegurar su calidad, consistencia y accesibilidad para la toma efectiva de decisiones y el cumplimiento de los objetivos estratégicos del negocio.

En un equipo de arquitectura de datos, se encuentran roles clave, como: el arquitecto de datos, que se encarga de diseñar e implementar la arquitectura alineada a los requerimientos del negocio; el administrador de datos, que gestiona y aplica las políticas, los estándares y los procesos de gestión de datos; el ingeniero de datos, que desarrolla y operacionaliza las soluciones de datos en las plataformas definidas por la arquitectura, y el analista de datos, que analiza los datos procesados y genera informes y visualizaciones.

Principios y mejores prácticas de la arquitectura de datos

Los modelos de datos son artefactos fundamentales en el diseño de la arquitectura de datos. El modelo conceptual representa una vista de alto nivel de las entidades y las relaciones de negocio relevantes. El modelo lógico define las estructuras de datos independientes de la plataforma tecnológica subyacente. Por último, el modelo físico especifica el esquema de datos en el motor de base de datos objetivo donde se implementará la solución.

Existen diversos estándares, como TOGAF, Zachman e ISO, y patrones de diseño, como el modelo Estrella, Copo de Nieve y el manejo de Datos Temporales, que guían la construcción de modelos de datos robustos, escalables y alineados con las mejores prácticas de la industria.

La gobernanza de datos es un marco integral de estrategias, métricas, roles, procesos y herramientas que permite ejercer control sobre los activos de datos corporativos. Algunos modelos populares de gobernanza son: el modelo centralizado, donde un equipo central gestiona y aplica las políticas de datos; el modelo descentralizado, en el que cada área funcional es responsable de sus propios datos, y el modelo híbrido, que combina aspectos centralizados y descentralizados. La democratización de datos busca facilitar el acceso y el uso de los datos por parte de diferentes usuarios dentro de la organización y fomenta una cultura orientada a los datos.

Finalmente, las políticas de seguridad y privacidad son cruciales para proteger los datos sensibles de la organización y garantizar el cumplimiento de las regulaciones y las normativas aplicables.

Ciclo de vida de la arquitectura de datos

El ciclo de vida de la arquitectura de datos cubre las principales fases de un proyecto de datos, comenzando con la adquisición y el entendimiento de las fuentes de datos internas y externas, así como la comprensión de los requerimientos del negocio. Luego, se aplican técnicas de limpieza y de preparación de datos para abordar problemas de calidad, como la detección de valores nulos o atípicos y la estandarización de formatos.

Una vez que los datos están listos, se procede al modelado y el diseño conceptual, lógico y físico según los estándares y los patrones establecidos en la arquitectura. Posteriormente, se implementan y despliegan las estructuras de datos en las plataformas definidas y se cargan los datos procesados.

Finalmente, se realiza el monitoreo y el mantenimiento continuo de la solución implementada, supervisando la calidad, el rendimiento y haciendo ajustes necesarios para garantizar el funcionamiento de la arquitectura de datos en el tiempo.

Herramientas y tecnologías para la arquitectura de datos

Para el modelado de datos se utilizan herramientas, como ER/Assistant que facilitan el diseño de modelos Entidad-Relación, Visual Paradigm para el modelado de datos UML y ErWin, una plataforma integral para el modelado de datos empresariales.

En cuanto al almacenamiento y el procesamiento de datos, los Data Warehouses, como Teradata y Netezza, son soluciones populares para almacenar datos estructurados. Los Data Lakes, como AWS S3 y Azure Data Lake, son repositorios para datos sin procesar. Además, existen motores analíticos como Spark y Presto, diseñados para el procesamiento eficiente de grandes volúmenes de datos (Big Data).

Para la gestión de la calidad y la gobernanza de datos, se utilizan herramientas como: Collibra, una plataforma integral de gobernanza de datos; Alation, que cataloga y gestiona los datos empresariales, y SAS DataGov, una solución completa para la calidad y la gestión de datos.

Finalmente, en el ámbito de la seguridad y la privacidad de datos, se encuentran herramientas como HashiCorp Vault, que gestiona y protege secretos y credenciales; AWS KMS, que crea y controla claves de cifrado de datos, e Informatica Secure@Source, una herramienta especializada en el enmascarado de datos sensibles.

Estas herramientas y tecnologías facilitan las tareas de diseño, implementación, mantenimiento y gobernanza de la arquitectura de datos corporativa y brindan soporte a lo largo de todo el ciclo de vida de los proyectos de datos.

Metodología

Este seminario está estructurado en tres bloques temáticos. El primero corresponde a la apertura del evento y a la reflexión sobre la importancia de la arquitectura de datos y de su articulación con la política pública a cargo de Jairo Alberto Riascos, de la Dirección de Gobierno Digital del Ministerio de Tecnologías de la Información y Comunicaciones en Colombia (MINTIC).

El segundo bloque corresponde a las presentaciones de cuatro expertos en arquitectura y ciencia de datos: *Introducción* a *la arquitectura de datos,* a cargo del experto Harvey Rosas, CTO Prevsis; *Principios y mejores prácticas de la arquitectura de datos*, liderado por el experto Álvaro Montenegro, profesor Universidad Nacional de Colombia; *Ciclo de vida de la arquitectura de datos,* impartido por el experto Tito Neira, Chief Data Strategy Officer en ADL Digital Lab; y *Herramientas y tecnologías para la arquitectura de datos*, a cargo de Vivian Aranda, Product Owner en ADL Digital Lab.

El tercer bloque abre un espacio para preguntas y algunas reflexiones finales a cargo de Jairo Alberto Riascos, de la Dirección de Gobierno Digital del Ministerio de Tecnologías de la Información y Comunicaciones en Colombia (MINTIC).

Conferencistas

Jairo Alberto Riascos Muñoz

Entusiasta de los datos y de las tecnologías emergentes con más de 15 años contribuyendo a iniciativas de arquitectura empresarial y transformación digital. Últimamente se ha dedicado a impulsar los datos como infraestructura en el país y la arquitectura empresarial en el sector público. Ha trabajado en la Contraloría General de la República (CGR) y en la Universidad de los Andes, y actualmente hace parte de la Dirección de Gobierno Digital, del Ministerio de Tecnologías de la Información y Comunicaciones en Colombia (MINTIC). Es ingeniero electrónico de la Universidad Nacional de Colombia, con una especialización en gerencia de tecnología de la Escuela de Administración de Negocios (EAN) y una maestría en arquitecturas de TI en la Universidad de los Andes.

Harvey Rosas

Destacado profesional en Matemática Aplicada, especializado en inteligencia artificial, minería de datos, análisis de texto automático, estadística computacional y educación. Obtuvo su PhD y Maestría en Matemáticas Aplicadas del Instituto Avanzado de Ciencia y Tecnología de Corea del Sur (KAIST) y su pregrado en Matemáticas en la Universidad Nacional de Colombia. Ha sido profesor en la Universidad de Valparaíso, impartiendo cursos en estadística computacional e inteligencia artificial y ha sido profesor visitante en diversas universidades. En el ámbito profesional, ha trabajado como CEO de un spin-off científico financiado por Colciencias en Colombia y actualmente es el CTO en Prevsis. Además, es miembro activo de Toastmasters

International y ha dado charlas inspiradoras, incluyendo una presentación en TEDx. Su investigación se centra en machine learning, minería de texto e inteligencia artificial aplicada y ha recibido el premio de Excelencia Docente de la Universidad de Valparaíso por su destacada labor educativa. Enlace a su charla TEDx.

Álvaro Mauricio Montenegro Díaz

Profesor Asociado de dedicación exclusiva en el Departamento de Estadística de la Universidad Nacional de Colombia, donde también dirige el grupo de investigación SICS. Es matemático y auditor de sistemas, con una Maestría y un Doctorado en Estadística de la Universidad Nacional de Colombia. Ha sido asesor de diversas instituciones estatales, incluyendo el Ministerio de Salud y Protección Social, el Ministerio de Ciencia y Tecnología, el Consejo Superior de la Judicatura y el Instituto Colombiano para la Evaluación de la Educación (ICFES). Ha desempeñado roles clave como director del Departamento de Estadística y director de Sistemas de Información en la Universidad Nacional. En el ámbito docente, imparte cursos en Minería de Datos, Aprendizaje Profundo, Ciencia de Datos, Big Data, Estadística Bayesiana y Teoría de Respuesta al Ítem, entre otros.

Tito Pablo Neira Ávila

Destacado especialista en estrategia de datos con más de 20 años de experiencia en el sector. Actualmente, se desempeña como Chief Data Strategy Officer en ADL Digital Lab, donde lidera la integración y la explotación de datos digitales y core para impulsar la transformación digital. Con un Doctorado en curso en Innovación de Datos, ha trabajado en roles clave en instituciones reconocidas como el Banco de Bogotá, Scotiabank Colpatria y El Tiempo Casa Editorial, donde ha desarrollado y ejecutado estrategias de datos y analítica. Además, es docente catedrático en la Universidad de Los Andes, donde imparte cursos en ciencia de datos y estrategia de datos. Su pasión por la toma de decisiones basada en datos, la creación de valor y la transformación digital lo han llevado a ser reconocido como uno de los 100 innovadores globales en datos y analítica. Es también un prolífico autor y conferencista en el ámbito de la ciencia de datos y la estadística.

Vivian Lucia Aranda Camacho

Vivian Lucia Aranda Camacho es una profesional destacada en el ámbito de la ingeniería de datos, arquitectura y gobernanza de datos. Posee una Maestría en Ingeniería de Información y una Especialización en Sistemas de Información de la Universidad de Los Andes. Con más de

una década de experiencia en el campo, ha ocupado roles significativos como ingeniera de datos y Product Owner en ADL Digital Lab, y arquitecta de datos en el Instituto Colombiano para la Evaluación de la Educación (ICFES). Su experiencia abarca la arquitectura de datos, ingeniería de big data e inteligencia de negocios, con un fuerte énfasis en la gestión de proyectos y el análisis de datos. Aranda también es AWS Certified Cloud Practitioner y ha sido reconocida por sus contribuciones a iniciativas de ciencia de datos, graduándose con honores del programa DS4A / Colombia 4.0.

Resultados de aprendizaje

Al finalizar este seminario, los asistentes estarán en capacidad de:

- Comprender los conceptos fundamentales de la arquitectura de datos y su importancia en el contexto de los proyectos de ciencia de datos.
- Identificar los principios, las mejores prácticas, los modelos, los estándares y los patrones clave para el diseño y la implementación efectiva de arquitecturas de datos.
- Conocer las diferentes etapas del ciclo de vida de la arquitectura de datos y los aspectos críticos a considerar en cada una de ellas.
- Reconocer las principales herramientas y tecnologías utilizadas en el modelado, el almacenamiento, el procesamiento, la calidad, la gobernanza y la seguridad de datos en una arquitectura de datos.
- Comprender la importancia de promover una cultura organizacional que respalde y facilite la adopción de prácticas sólidas de arquitectura de datos.
- Adquirir conocimientos generales sobre las mejores prácticas para el diseño, la implementación y el mantenimiento de arquitecturas de datos robustas.
- Apreciar la importancia de la colaboración efectiva entre los diferentes roles involucrados en la arquitectura de datos.
- Reconocer la necesidad de integrar la arquitectura de datos con los desarrollos operacionales para mejorar las interfaces y la usabilidad.

Agenda

Tiempo	Actividad	Responsable
8:00 a. m. a 8:30 a. m.	Instalación y registro de asistentes.	Equipo DIRPEN.
8:30 a. m. a 8:45 a. m.	Apertura del evento y foto oficial.	Lina Casas Quiroga, Dirección de Difusión y Cultura Estadística (DICE).
8:45 a. m. a 9:00 a. m.	Presentación de la agenda y del perfil de los conferencistas.	
9:00 a. m. a 9:15 a. m.	Palabras de bienvenida.	Andrea Ramírez Pisco, subdirectora DANE Julieth Solano Villa, directora técnica DIRPEN.
9:15 a. m. a 9:30 a. m.	Importancia de la arquitectura de los datos y su articulación con la política pública.	Jairo Alberto Riascos Muñoz, MINTIC.
9:30 a. m. a 10:00 a. m.	*Introducción a la arquitectura de datos*: • Definición y conceptos clave. • Importancia de la arquitectura de datos en proyectos de ciencia de datos. • Roles y responsabilidades en un equipo de arquitectura de datos.	Harvey Rosas, CTO Prevsis.
10:00 a. m. a 10:10 a. m.	Espacio de preguntas frente a la conferencia "*Introducción a la arquitectura de datos*".	Modera: Lina Casas Quiroga, DICE. Respuestas: Harvey Rosas, CTO Prevsis.
10:10 a. m. a 10:40 a. m.	*Principios y mejores prácticas de la arquitectura de datos*: • Modelos de datos (conceptual, lógico y físico). • Estándares y patrones de diseño. • Gobernanza y calidad de datos. • Modelos de gobernanza de datos: fortalezas y desafíos. • Democratización de datos y su importancia en la gobernanza de estos. • Seguridad y privacidad de datos.	Álvaro Montenegro, profesor Universidad Nacional de Colombia.

Tiempo	Actividad	Responsable
10:40 a. m. a 10:50 a. m.	Espacio de preguntas frente a la conferencia "*Principios y mejores prácticas de la arquitectura de datos*".	Modera: Lina Casas Quiroga, DICE. Respuestas: Álvaro Montenegro, profesor Universidad Nacional de Colombia.
10:50 a. m. a 11:00 a. m.	Café	
11:00 a. m. a 11:30 a. m.	*Ciclo de vida de la arquitectura de datos*: • Adquisición y entendimiento de los datos. • Limpieza y preparación de datos. • Modelado de datos. • Implementación y despliegue. • Monitoreo y mantenimiento.	Tito Neira, ADL Digital Lab.
11:30 a. m. a 11:40 a. m.	Espacio de preguntas frente a la conferencia "*Ciclo de vida de la arquitectura de datos*".	Modera: Lina Casas Quiroga, DICE. Respuestas: Tito Neira, ADL Digital Lab.
11:40 a. m. a 12:10 m.	*Herramientas y tecnologías para la arquitectura de datos*: • Herramientas de modelado de datos. • Plataformas de almacenamiento y procesamiento de datos • Herramientas de calidad y gobernanza de datos. • Herramientas de seguridad y privacidad de datos.	Vivian Aranda, ADL Digital Lab.
12:10 m. a 12:20 m.	Espacio de preguntas frente a la conferencia "*Herramientas y tecnologías para la arquitectura de datos*".	Modera: Lina Casas Quiroga, DICE. Respuestas: Vivian Aranda, ADL Digital Lab.
12:20 m. – 12:30 m.	Conclusiones y cierre.	Jairo Alberto Riascos Muñoz, MINTIC.

Referencias

- Departamento Administrativo Nacional de Estadística. (2023). *Plan Estadístico Nacional (2023-2027)*.
- Departamento Nacional de Planeación. (2022). *Plan Nacional de Desarrollo 2022-2026. ¨Colombia Potencia mundial de la vida¨*.
- Fowler, M. (2003). *Patterns of Enterprise Application Architecture*. Addison-Wesley Professional.
- Inmon, W. H., & Linstedt, D. (2016). *Data Architecture: A Primer for the Data Scientist*. Morgan Kaufmann.
- Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley.
- Ministerio de Tecnología de la Información y las Comunicaciones en Colombia, Departamento Nacional de Planeación y Presidencia de la República de Colombia. (2021). *Plan Nacional de Infraestructura de Datos*.
- Redman, T. C. (2008). *Data Driven: Profiting from Your Most Important Business Asset*. Harvard Business Press.