Consejo Asesor Técnico del Sistema Estadístico Nacional (CASEN) 2023 - 2025
ACTA TERCERA SESIÓN ORDINARIA - AYUDA DE MEMORIA

Ciudad: Bogotá D.C
Lugar: Sesión virtual
Tema: Modelo Tecnologico para Produccion de registros estadisticos base de unidades economicas y poblacion

Hora: 02:00 p. m. a 4:00 p. m.
Fecha: 15/08/2025
Dependencia responsable: OSIS -DIRPEN

Participantes

Miembros de la Sala Especializada para la Modernización Tecnológica de la Producción Estadística del CASEN

Mario Linares Vásquez.
Nicolás Cardozo Álvarez.
León Darío Parra.

Departamento Administrativo Nacional de Estadística (DANE)

Diego Antonio Campos Cáceres, Asesor OSIS

Andrea Catherine Neira Bustamante, designada OSIS

Guilbert Leonardo Reina Colorado, designado OSIS
Juan Manuel Rivera Cabezas, designado OSIS

Lady Gilari Torres Becerra, designada OSIS

Mario Silva Leal, designado OSIS

Yohan Ricardo Cespedes Villar, designado DRE

Secretaría Técnica del CASEN - DIRPEN

Maria del Pilar Gómez Arciniegas, directora técnica
DIRPEN (E)

Germán Leonidas Orjuela Borda, coordinador
GIT Planificación y Articulación Estadística, DIRPEN.

Elizabeth Moreno Barbosa, asesora.

Derly Vivian Lizarazo García, responsable
Sala Especializada para la Modernización
Tecnológica de la Producción Estadística.

AGENDA

Tiempo	Actividad	Responsable
2:00 p.m. a 2:05 p. m.	Instalación, verificación del quorum y registro fotográfico.	Derly Lizarazo, responsable de la sala.
2:05 p. m. a 2:10 p. m.	Apertura de la reunión	Germán Leonidas Orjuela, coordinador GIT PAE
2:10 p. m. a 2:15 p. m.	Síntesis reunión anterior	Derly Lizarazo, responsable de la sala.
2:15 p. m. a 3:00 p. m.	Presentación modelo tecnológico que se implementará para la producción de registros estadísticos base de unidades económicas y población	Yohan Céspedes, registros estadísticos
3:00 p. m. a 3:50 p. m.	Realimentación por parte de los miembros de la sala	Miembros de la Sala Modera: Derly Lizarazo
3:50 p. m a 3:55 p. m.	Compromisos	Derly Lizarazo, responsable de la sala.
3:55 p. m a 4:00 p. m.	Conclusiones y cierre.	Germán Leonidas Orjuela, coordinador GIT PAE

Desarrollo

Objetivo

Presentar el modelo tecnológico que se implementará para la producción de registros estadísticos base de unidades económicas y población para recibir realimentación por parte de los expertos.

1. Apertura

Germán Leonidas Orjuela Borda, coordinador del GIT Planificación y Articulación Estadística, DIRPEN, realizó la apertura de la reunión, destacando el propósito principal de este espacio el cual fue presentar el modelo tecnológico que se implementará para la producción de registros estadísticos base de unidades económicas y población para recibir retroalimentación del equipo.

2. Síntesis de la reunión anterior

Derly Lizarazo del GIT Planificación y Articulación Estadística, desarrollo la síntesis, destacando que en el espacio previo tuvo como propósito presentar y socializar el procedimiento de ingreso seguro a los sistemas de información del DANE, utilizando el esquema Single Sign-On (SSO). Este mecanismo busca fortalecer la seguridad de la información institucional, facilitando el acceso controlado y centralizado a los sistemas. Además, se buscó recoger observaciones y recomendaciones de los participantes para mejorar y robustecer el protocolo propuesto.

3. Presentación modelo tecnológico que se implementará para la producción de registros estadísticos base de unidades económicas y población

Yohan Céspedes de la Dirección de Registros Estadísticos, realizó la presentación en la cual se destacó:

¿Qué es SIREVE?

SIREVE (Sistema de Integración de Registros Estadísticos Base) es una solución tecnológica desarrollada por el DANE para transformar registros administrativos en registros estadísticos confiables, estandarizados y explotables. Su propósito es consolidar información de múltiples fuentes en una estructura común que permita análisis estadísticos robustos y seguros.

Componentes Fundamentales del Modelo

1. Estructura por Capas

SIREVE se organiza en tres capas principales:

- Raw: Datos crudos, seudonimizados para proteger la identidad de personas y empresas.
- Staging: Datos validados y normalizados, listos para ser integrados en estructuras multidimensionales.
- Datamart (BI): Datos procesados para análisis estadístico, visualización y explotación por usuarios finales.

2. Procesos ETL

El sistema realiza procesos de Extracción, Transformación y Carga (ETL) en dos fases:

- ETL1: Seudonimización y validación inicial.
- ETL2: Integración y estructuración en registros base y satélites.

Herramientas Tecnológicas Utilizadas

- Docker: Para contenerización y despliegue escalable.
- Python + PySpark: Para procesamiento distribuido.
- Pydantic: Validación de modelos de datos.
- FastAPI: Servicios para pseudonimización y generación de ID estadísticos.
- Apache Airflow: Orquestación de procesos ETL.
- Oracle: Motor de base de datos principal.

Seguridad y Gobernanza

- Pseudonimización: Cada entidad recibe un ID estadístico único sin relación directa con datos sensibles.
- Control de acceso basado en roles (RBAC): Diferentes niveles de usuarios (configurador, ingeniero, especializado, general, superusuario).
- Logs auditables: Registro de todas las operaciones para trazabilidad y auditoría.

ID Estadístico

- Compuesto por un prefijo (01 para personas, 02 para empresas) + código hexadecimal incremental.
- Permite identificar entidades sin exponer datos sensibles.
- Se genera a partir de la primera fuente de información disponible.

Data Linkage

- Estrategia en cascada para vincular registros que podrían pertenecer a la misma entidad.
- Uso de la librería Splink con metodología Fellegi-Sunter para comparar atributos como nombre, fecha de nacimiento, ubicación, etc.
- Evita duplicidades y errores en la asignación de ID.

Interoperabilidad y Escalabilidad

- Capacidad para integrar múltiples fuentes con distintos formatos (APIs, archivos planos, vistas, motores de BD).
- Arquitectura modular que permite crecimiento según demanda.
- Planeación para incluir visualizadores como Power BI, Tableau, o herramientas libres.

Impacto Esperado

- Unificación de fuentes: Permite análisis consistentes y comparables.
- Automatización de procesos: Reducción de tiempos y errores.
- Trazabilidad y monitoreo: Seguimiento de actualizaciones y uso de datos.
- Escalabilidad: Posibilidad de incluir nuevas dimensiones como inmuebles y actividades.

Estado Actual y Próximos Pasos

- El sistema está en fase de implementación con pruebas internas.
- Se han configurado servicios para generación de ID estadísticos.
- Próximos pasos incluyen:
    - Inclusión de nuevas fuentes.
    - Publicación de dashboards.

- Implementación de APIs de consulta.
- Escalado a entorno productivo.
- Gestión de metadatos y datos maestros.

4. Realimentación por parte de los miembros de la sala

Frente a lo presentado los miembros de la sala emitieron sus recomendaciones.

Mario Linares Vásquez

- Sugirió incluir una capa de observabilidad y monitoreo para detectar errores en tiempo real durante el procesamiento de datos.
- Propuso el uso de catálogos indexados y estructuras de búsqueda como hash tables y búsqueda binaria para optimizar el proceso de data linkage.
- Recomendó explorar el uso de GPU para procesamiento paralelo de grandes volúmenes de datos.
- Planteó la idea de distribuir catálogos en múltiples bases de datos (sharding) para evitar cuellos de botella.
- Señaló la importancia de microoptimizar el código, especialmente en operaciones de comparación de cadenas.

León Darío Parra Bernal

- Preguntó por la estructura de gobernanza de datos, incluyendo validación de controles y auditoría.
- Recomendó realizar pruebas funcionales con usuarios externos para validar la experiencia y utilidad del sistema.
- Sugirió explorar algoritmos de matching probabilístico como Jaccard y Propensity Score Matching.
- Propuso considerar la virtualización de datos como estrategia para mejorar eficiencia y reducir costos.

Nicolás Cardozo Álvarez

- Consultó sobre el momento en que se realiza la validación de datos en el flujo de ingreso.
- Sugirió implementar estrategias de búsqueda como Bicon Guide Search, que segmentan datos en subconjuntos para facilitar búsquedas eficientes.

5. Conclusiones y Cierre

Derly Lizarazo agradeció a todos los participantes por sus aportes y confirmó los compromisos establecidos:

- Envío del acta para revisión y aprobación.
- Remisión de preguntas orientadoras a los profesores.
- Envío del material y asunto de consulta para la próxima sesión, programada para el **29 de agosto**, donde el equipo de OSIS presentará los ajustes al procedimiento discutido en la reunión anterior.

Se destacó que los comentarios de los profesores sobre el modelo tecnológico SIREVE como la inclusión de una capa de observabilidad, el uso de virtualización y algoritmos de matching, serán tenidos en cuenta para mejorar el sistema. Se acordó realizar una nueva presentación del modelo ajustado en una reunión futura.

Finalmente, **María del Pilar Gómez Arciniegas** agradeció la participación y resaltó la importancia del trabajo realizado para fortalecer el uso de registros administrativos como fuente de información estadística.

Compromisos

Tarea	Envío del acta para revisión y aprobación.
Responsable	Derly Lizarazo, responsable de la sala.
Fecha entrega	20/08/2025
Tarea	Remisión de materiales y asunto de consulta para la próxima sesión.
Responsable	Derly Lizarazo, responsable de la sala.
Fecha entrega	22/08/2025

Próxima reunión:

Responsable de convocar: DIRPEN
Fecha: 29 de agosto de 2:00 a 4:00 p.m