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

Ciudad: Bogotá D.C
Lugar: Sesión virtual
Tema: Desarrollo de sistemas de información

Hora: 02:00 p. m. a 4:00 p. m.
Fecha: 07/11/2025
Dependencia responsable: DIRPEN

Participantes

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

Nicolás Cardozo Álvarez.
León Darío Parra.

Departamento Administrativo Nacional de Estadística (DANE)

Diego Antonio Campos Cáceres, asesor OSIS

Andrea Catherine Neira Bustamante, designada OSIS

Pedro Antonio Rubio, designado Calidad Estadística

Marisol Sabogal Hoyos, designada Calidad Estadística

Maria del Pilar Gómez Arciniegas, Coordinadora GIT Calidad Estadística

Secretaría Técnica del CASEN – DIRPEN

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	Derly Lizarazo, responsable de la sala.
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 del modelo actual de evaluación de la calidad de archivos de datos de operaciones estadísticas	Pedro Rubio, GIT Calidad Estadística
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.	Derly Lizarazo, responsable de la sala.

Desarrollo

Objetivo

Exponer el modelo actual de evaluación de la calidad de base de datos de operaciones estadísticas, con el propósito de recibir observaciones y aportes de expertos sobre sus componentes.

1. Apertura

Derly Lizarazo, realizó la apertura de la reunión, destacando el propósito principal de este espacio. Se confirmó asistencia de los expertos Leon Darío Parra y Nicolás Cardozo, como delegados de OSIS y DIRPEN.

2. Síntesis de la reunión anterior

Derly Lizarazo del GIT Planificación y Articulación Estadística, desarrollo la síntesis, destacando que en el espacio previo se socializó la guía para la construcción de un sistema de información estadística, con el propósito de recoger observaciones y recomendaciones por parte de expertos, que contribuyan a su fortalecimiento y mejora.

3. Presentación del modelo actual de evaluación de la calidad de archivos de datos de operaciones estadísticas

Pedro Rubio del GIT Calidad Estadística realizó la presentación del tema en los siguientes aspectos:

Objetivo del modelo:

Garantizar que las bases de datos utilizadas en operaciones estadísticas cumplan con estándares de calidad definidos por la Norma Técnica NTC PE 1000:2020 y la Ley 2335 de 2023 para atributos de calidad de datos (precisión, completitud, consistencia, exactitud, conformidad).
Asegurar información confiable, precisa y estandarizada para la toma de decisiones y la producción estadística nacional.

Preparación de insumos:

Archivos de datos (bases/tablas).
Formato de identificación de evidencias (diccionario de datos con características de cada variable).
Archivo de traducción informática para reglas de validación.

Carga y limpieza de datos:

Ajuste de variables (mayúsculas, longitud ≤ 26 caracteres, eliminación de caracteres especiales).
División de bases con más de 1 millón de registros.
Importación en Oracle SQL Developer.

Creación y ejecución de reglas de validación:

Reglas en SQL para verificar atributos (no nulos, rangos válidos, formatos correctos).
Validación manual y pruebas en Oracle antes del procesamiento.
Codificación de reglas según dimensión afectada (completitud, exactitud, etc.).

Procesamiento automatizado:

Uso de Pentaho Data Integration v7.1 para aplicar reglas y generar resultados.
Procesos: carga de reglas, evaluación de calidad, manejo de inconsistencias.
Duración variable (1 a 12 horas según volumen y errores).

Resultados y reportes:

Archivos en Excel: resultados por registro, campo y variable y resumen de inconsistencias (para revisión por la entidad).

Indicadores calculados: registros procesados vs inconsistentes, campos analizados vs afectados y variables afectadas (%).

Dimensiones evaluadas: consistencia, completitud, actualidad, exactitud, precisión, conformidad.

Comunicación y validación: Envío del resumen de inconsistencias a la entidad (3 días hábiles para respuesta), ajustes según retroalimentación e informe final con conclusiones, hallazgos y oportunidades de mejora (5 días hábiles).

Herramientas utilizadas: Oracle SQL Developer (carga y validación), Pentaho Data Integration (procesamiento) y Excel (reportes e indicadores).

Marisol Sabogal oderó las preguntas orientadoras para ser respondidas por los expertos.

4. Retroalimentación por parte de los miembros de la sala

Frente a las preguntas orientadoras los profesores realizaron sus comentarios y recomendaciones:

Nicolás Cardozo Álvarez

- Observó que el proceso actual es altamente manual, lo que genera ineficiencias y riesgos de error.

- Propuso automatizar la validación y generación de reportes directamente sobre la base de datos, evitando la dependencia de archivos intermedios en Excel y reduciendo pasos redundantes.
- Sugirió explorar herramientas y metodologías modernas para análisis de calidad de datos:
  - CDQ (Comprehensive Data Quality) como marco metodológico.
  - Lenguajes especializados como DataLog, que permiten definir reglas lógicas y validarlas sobre conjuntos de datos de manera eficiente.
- Recomendó investigar soluciones que soporten procesamiento paralelo para grandes volúmenes de datos, mencionando tecnologías como:
  - Apache Hadoop y su ecosistema.
  - Apache Spark (incluyendo librerías para calidad de datos).
  - Herramientas basadas en Databricks y frameworks en Python como DQX para análisis de calidad.
- Enfatizó que estas alternativas podrían mejorar la eficiencia y escalabilidad, reduciendo tiempos de ejecución y optimizando recursos.

León Darío Parra Bernal

- Coincidió en que Pentaho está rezagado frente a herramientas modernas para Big Data, y que su uso prolongado limita el rendimiento y la capacidad de respuesta.
- Recomendó migrar hacia plataformas más robustas como:
  - Apache Spark para procesamiento distribuido.
  - Servicios en la nube como AWS (Amazon Web Services) para escalabilidad y gestión eficiente de grandes volúmenes.
- Señaló que, si no es posible migrar de inmediato, se debe aplicar partición de datos para evitar represamientos y reducir riesgos en procesos largos (que pueden tardar más de 12 horas).
- Sugirió incorporar buenas prácticas de gobernanza de datos, incluyendo:
  - Indicadores de seguimiento por nivel (estratégico, táctico y operativo).
  - Matrices de riesgo y semáforos para monitorear problemas en tiempo real.
  - Protocolos de seguridad de la información alineados con estándares internacionales.
- Indicó que la comunicación de resultados debe ser más visual y narrativa, orientada a tomadores de decisiones, mostrando impacto en costos, eficiencia y riesgos.
- Recalcó que mantener procesos manuales incrementa riesgos de sesgos y errores, por lo que se deben aplicar controles para mitigarlos.

Compromisos

Tarea	Envío del acta para revisión y aprobación.
Responsable	Derly Lizarazo, responsable de la sala.
Fecha entrega	11/11/2025

Próxima reunión:

Responsable de convocar: DIRPEN
Fecha: 28 de noviembre de 2:00 a 4:00 p.m