Proyecto 2: Infraestructura Visible

AUTORES

Samuel Nicolas Rodriguez Celis

Edward David Diaz Fontecha

Andres Leonardo Bayona Latorre

Universidad de los Andes

INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

ISIS 3301: Inteligencia de Negocios

Bogotá D.C.

05 de diciembre de 2022

Tabla de contenido

1. Identificación de necesidades analíticas

Tema analítico	Análisis requeridos	Categoría del análisis	Procesos de negocio	Fuentes de datos y datos
Rol de la minería en el desarrollo del conflicto armado	Relación entre número de casos de violencia y magnitud de la actividad minera	Tablero de control	Información de desplazados dada por el gobierno y análisis de proyectos por entidad	- Unidad para la Atención y Reparación Integral a las Víctimas - Proyectos mineros
	Percepción de la seguridad en los habitantes de zonas con alta presencia minera	Minería de datos	Encuestas y análisis de proyectos por entidad	- Percepción – CIUDATOS - Proyectos mineros
	Comportamiento de las tasas de homicidio según el tipo y el departamento	Análisis OLAP	Datos de medicina legal y análisis de proyectos por entidad	- Homicidios – INMLCF - Proyectos mineros
Impacto de la minería sobre la	Probabilidad de aparición de ciertas enfermedades por	Minería de datos	Estudio sobre enfermedades más	- Presencia de ciertas

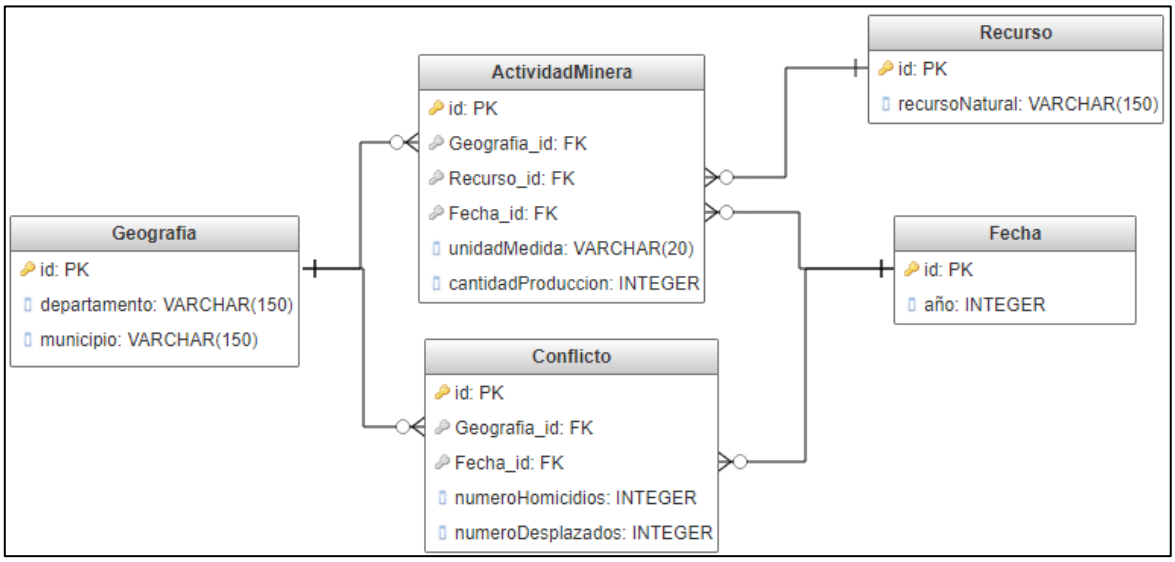
salud de los habitantes	presencia de minería		relacionadas con la minería y análisis de proyectos por entidad	enfermedades en el país – MSPS - Proyectos mineros
	Cobertura del estado (vacunas & PPNA) y satisfacción de las personas con la salud en zonas de alta actividad minera	Tableros de control	Encuestas, análisis de proyectos mineros en las entidades con mayor abandono del estado	- Satisfacción – CIUDATOS - Cobertura y presencia del estado – MSPS

- http://www.scielo.org.co/pdf/crim/v58n1/v58n1a04.pdf
- https://alacip.org/cong13/695-angel-7cc.pdf
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252433
- https://www.cdc.gov/spanish/niosh/mining/topics/enfermedades.html

Como los anteriores encontramos muchos artículos y proyectos en los que se han encontrado relaciones relevantes entre la alta presencia de actividades minera y la aparición de ciertas enfermedades (más que todo respiratorias) y una alta tasa de violencia. A primera vista parece ser que estos son problemas que son más comunes en departamentos y municipios con poca presencia del estado y que, por ende, son más pobres o son más vulnerables ante ataques de grupos armados o delincuencia organizada. Además, consideramos importante el hecho de que el recurso natural también ayuda a determinar la magnitud de estos problemas; en algunos casos se hablaba de que recursos como el carbón atraen más a estas bandas criminales y por ende será de utilidad realizar un análisis no solo sobre los departamentos o municipios sino también sobre el recurso natural extraído en cada proyecto minero.

2. Modelar Data Marts

2.1. Modelos multidimensionales



2.2. Justificación del modelo

2.2.1. Granularidad

En cuanto a granularidad decidimos que en la fecha se iba a manejar únicamente el año, puesto que en las fuentes a pesar de que existía el concepto de trimestre rara vez era diferente del trimestre 4, es decir, la mayoría de los datos eran tomados una única vez por año.

Para el caso de la geografía decidimos ser un poco más precisos, incluyendo así departamento y municipio de manera que los análisis puedan hacerse desde varias perspectivas, la idea es entonces que el usuario pueda finalmente filtrar la información de las gráficas según el detalle geográfico que desee.

2.2.2. Hechos y medidas

- ActividadMinera (tabla de hechos): La única medida en este caso es la cantidad de producción que se tendrá por cada recurso. Esta será aditiva y nos servirá para determinar el departamento o municipio con mayor cantidad extraída acumulada.
- Conflicto (tabla de hechos): Tendremos 2 medidas, haciendo referencia a la cantidad de homicidios y el número de desplazados. Ambas al representar una cantidad serán aditivas y la idea es que se pueda mostrar un acumulado en los tableros de control.

2.2.3. Atributos

En nuestro modelo los atributos de las dimensiones contienen información que en principio no debería cambiar, y en caso de hacerlo no nos interesa mantener una historia de estos cambios, pues son reportes verificados, entonces se usará el tipo 1 de manejo de historia.

3. Entendimiento, Data Mart y ETL

3.1. Entendimiento de las fuentes de datos

3.1.1. Conflicto armado

- Datos generales:

Se tienen un total de 314119 registros, los cuales tienen valores para 12 atributos diferentes.

- Posibles aplicaciones:

Tenemos información relacionada al número de personas secuestradas o desplazadas, los cuales son indicadores comunes al momento de analizar el impacto de guerras por el control minero en ciertas zonas delicadas del país.
Al igual que en las otras fuentes de datos, tenemos atributos referentes a fechas y ubicaciones, de manera que se facilita mucho la integración de información y se pueden hacer análisis que muestren el cambio de ciertos factores en el espacio y en el tiempo.

- Calidad de los datos:
  - ✓ Unicidad

    100% - no tenemos registros duplicados
  - ✓ Completitud

    Dato Cualitativo ≈ 1.5% - Dato Numérico ≈ 82.5% - Departamento, Código entidad, … ≈ 100%

    Encontramos que hay muy pocos registros con valores en “Dato Cualitativo”(ni el 1.6% del total). En cuanto a “Dato Numérico” tenemos una completitud aceptable, por encima del 80%. Para las demás columnas tenemos más del 99.9996% de la completitud.
  - ✓ Consistencia

    100% - no tenemos valores inconsistentes con lo especificado en el glosario. Se esperaba que en atributos como “Departamento” o “Mes” existieran inconsistencias (es común en otras fuentes), sin embargo, todo está en orden.
  - ✓ Validez

    Encontramos un registro que tenía varios atributos vacíos y que presentaba valores anormales en los atributos de “Indicador” y “Fuente”, los cuales no podían ser asociados a ninguno de los descritos en el glosario. Es por esto por lo que se decide eliminarlo definitivamente.

3.1.2. Salud

- Datos generales:

Se tienen un total de 411259 registros, los cuales tienen valores para 12 atributos diferentes.

- Posibles aplicaciones:

Tenemos más que todo tasas de mortalidad o de incidencia de diversas enfermedades, las cuales nos pueden servir para determinar en qué zonas del país hay más presencia de ciertas enfermedades o incluso en cuales el servicio de salud es más ineficiente al momento de combatirlas.
Hay ciertos indicadores que nos pueden ayudar a determinar el abandono del estado que existe en un municipio o entidad en específico, puesto que un análisis

apropiado sobre datos como la cobertura de vacunación o la población pobre no atendida (PPNA) puede servir a este propósito.
Como en las demás fuentes también tenemos datos geográficos y de tiempo, de manera, que se podrán relacionar de manera adecuada al momento de crear los tableros de control y realizar los análisis respectivos.

- Calidad de los datos:
  - ✓ Unicidad

  ≈ 100% - Se tienen 0 registros duplicados
  - ✓ Completitud

  Dato Cualitativo≈ 0% - Dato Numérico ≈ 80,72 %

  Es curioso que la columna de dato cualitativo no tiene ningún valor a lo largo de todos los registros, por lo cual es una columna que se puede ignorar. En segundo lugar, está el dato numérico que, si bien tiene datos, casi un 20% de los registros no cuentan con este valor, por lo que no nos serán de utilidad. En cuanto a los demás, no todos tienen 100% de completitud, lo cual sucede porque hay algunos registros que no tienen valores en los mismos atributos.
  - ✓ Consistencia

  A diferencia de la anterior fuente, en este caso no hay problemas de consistencia ni con los indicadores, ni con los departamentos o entidades.
  - ✓ Validez

  No encontramos errores de validez, todos los valores están dentro de los límites establecidos y las descripciones dadas en el glosario.

3.1.3. Minería

- Datos generales:

Se tienen un total de 45215 registros, los cuales tienen valores para 11 atributos diferentes.

- Posibles aplicaciones:

Atributos como la cantidad de producción o los valores de las contraprestaciones nos pueden dar una idea de cuáles son los departamentos o municipios con mayor actividad minera del país. Además, podríamos determinar la cantidad de proyectos realizados por zona para definir dónde hay mayor presencia de diferentes instituciones o grupos mineros en el país.
Tenemos datos de geografía y tiempo para cada registro, lo cual permitiría hacer una integración apropiada con las otras fuentes de información para hacer análisis de variación con respecto al tiempo o el lugar. Aunque la forma de representar los datos no es la misma para todas las fuentes (meses/trimestres, por ejemplo), es posible hacer ciertas transformaciones de cara a interpretar y analizar la correlación entre los datos.
Además de tener análisis por cantidad de proyectos o intensidad de la actividad minera, también podríamos clasificar según los recursos extraídos, ya que como veremos más adelante es posible relacionar la extracción de ciertos materiales con ciertos factores de salud o violencia.

- Calidad de los datos:
  - ✓ Unicidad

  ≈ 99. 97% - Se tienen 5 registros duplicados que se pueden eliminar

✓ Completitud

cantidad_producci_n ≈ 99.9977 % - valor_contraprestacion ≈ 99.8673 %

Se tiene una completitud de los datos bastante positiva ya que son pocos los registros que tienen algún valor nulo y únicamente 2 atributos de los 11 presentes cuentan con algunos vacíos. Por ende, es válido mantenerlos.

✓ Consistencia

Para el atributo “departamento” encontramos que ‘Nariño’ tenía dos representaciones, una correcta (la mayoría de los registros la tenían) pero otra con caracteres desconocidos reemplazando la ñ. En este caso es necesario corregirla para que esté correctamente representada.

✓ Validez

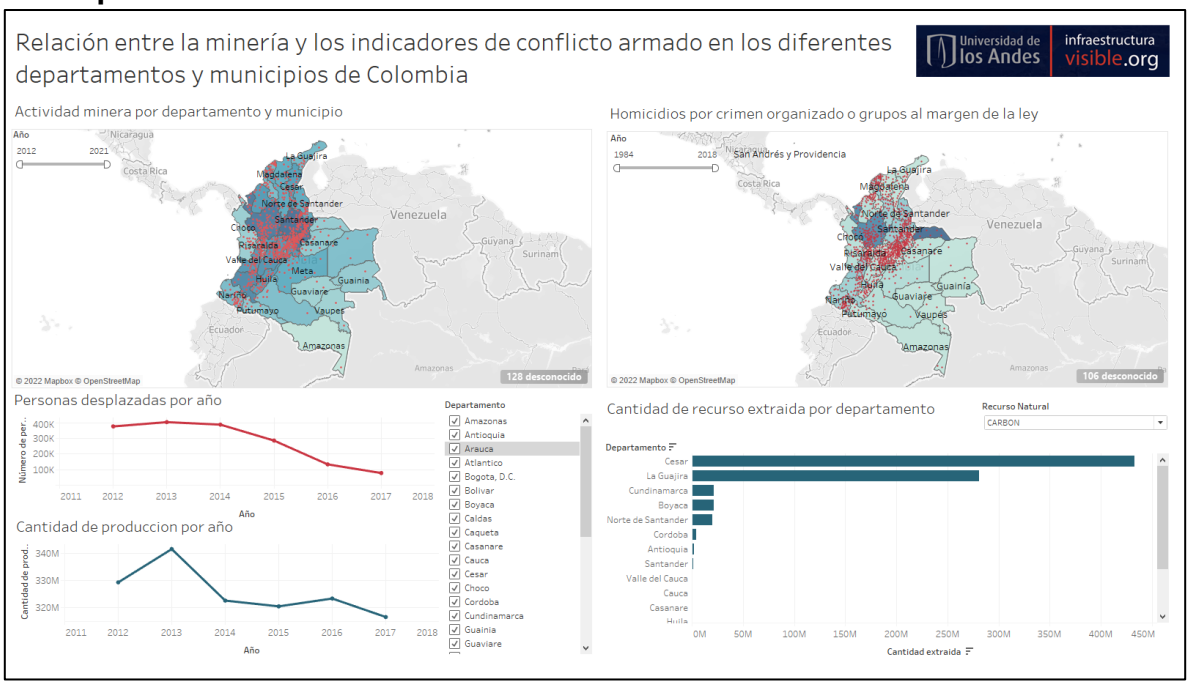
No encontramos errores de validez, todos los valores están dentro de los límites establecidos y las descripciones dadas en el glosario.

3.2. Diseño e implementación del proceso de ETL

4. Arquitectura de solución y tableros de control

4.1. Arquitectura de solución

4.2. Implementación de tableros de control



5. Distribución de actividades

Actividad	Responsable(s)
Necesidades analíticas: entrevista, investigación de proyectos similares	Nicolas Rodriguez
Elaboración y justificación de los modelos dimensionales	Nicolas Rodriguez & Andrés Bayona
Entendimiento de las fuentes de datos	Nicolas Rodriguez
Limpieza y preprocesamiento de los datos	David Diaz

Diseño e implementación del proceso de ETL	Andrés Bayona & David Diaz
Implementación tablero de conflicto armado	Nicolas Rodriguez
Realización de la presentación y el video	Todos

Integrante	**Puntaje /100**
Andrés Bayona	33
David Diaz	33
Nicolas Rodriguez	33