Sistema de análisis de datos sobre las compras estatales (caso Chile) bajo Big Data

Francisco Rossi, Iván Monjardin
Docente: Pedro Bonillo

Universidad ORT Uruguay
Ingeniería en Sistemas - Herramientas de Software para Big Data
Presentación de defensa del Obligatorio
Montevideo, Uruguay 2022

Resumen

En la actualidad, varios gobiernos disponibilizan datos sobre las compras realizadas por las instituciones públicas. El análisis de dichos datos puede dar mayor transparencia sobre el destino de fondos públicos. Sin embargo realizar el análisis puede ser costoso y requerir de un trabajo previo antes de poder resolver las consultas deseadas.

Este proyecto consiste en realizar dicho trabajo para análisis consultas sobre el caso particular del Estado chileno.

Palabras claves: Jupyter Notebooks, Apache Spark, Compras, Transparencia, Chile

Objetivos

General

Utilizar gráficas y otros elementos visuales para mostrar resúmenes sobre las compras del estado chileno. Realizar un análisis de los gastos públicos para formular y responder consultas que puedan ser de interés.

Especfíficos

1. Definir arquitectura y herramientas a utilizar para resolver el objetivo general
2. Realizar ingesta y unión de los datos de la página del Estado chileno
3. Realizar ingeniería de atributos
4. Analizar los datos y responder preguntas de investigación
5. Utilizar Solr como motor de búsqueda.

Metodología

- El trabajo se basó en metodologías y principios de desarrollo ágiles.
- El proyecto comenzó el 11/5, duró 4 semanas
- El docente de la materia tomó el rol de Product Owner con el cual tuvimos reuniones semanales para guiarnos sobre las tareas realizadas en función de nuestro objetivo.
- Los entregables del trabajo son:
    - Informe de trabajo realizado con documentación de ingeniería de atributos realizado
    - Presentación Power Point para la defensa
    - Jupyter Notebook y cualquier otro artefacto de código o similar que desarrollamos
    - Dataset depurado en formato csv

Problema

El problema consiste en mostrar de manera intuitiva y transparente los datos de las compras realizadas por el Estado chileno. Actualmente muchas preguntas que se pueden realizar los ciudadanos no son necesariamente triviales a responder. Aunque la información dada por el gobierno, en la página ChileCompra, puede ser completa, realizar consultas sobre ella no siempre es fácil. Se requiere un procesamiento y análisis de más de un millón de registros, para poder visualizar y realizar consultas sobre los datos.

Preguntas de investigación

- ¿Cuál es el gasto total del estado chileno en los meses analizados?
- ¿En qué áreas está el Estado chileno invirtiendo más? (Por sector y por rubro)
- ¿Cuántos gastos mayores al millón de dólares fueron efectuados entre noviembre de 2021 y mayo de 2022?
- ¿En qué ciudades/regiones se realizaron los mayores gastos?
- ¿Cuál fue el gasto promedio de una compra por mes del estado?
- ¿Cuál fue el gasto promedio en general?
- ¿Cuál fue la compra más repetida del estado?
- ¿Qué proyecciones se pueden hacer a futuro sobre los datos analizados?
- ¿Cuál es el tiempo promedio desde que se solicita la orden de compra, hasta que se autorice?
- ¿Cómo se distribuyen las compras en función de organismos?

Justificación y alcance

Justificación

El fin del proyecto es responder consultas relevantes sobre grandes volúmenes de datos utilizando herramientas y arquitecturas de Big Data.

Alcance

Realizar un informe con gráficas y resúmenes interesantes sobre los datos, respondiendo preguntas que se plantearon sobre los mismos. Todas las consultas se realizan sobre datos del periodo Nov-2021 y Mayo-2022 (2.5 millones de registros aprox.)

Limitaciones

- Tiempo: Para desarrollar el proyecto previo a la fecha de entrega y para trabajar en conjunto, tomando en cuenta otras obligaciones como trabajo o estudio de otras materias.
- Datos con errores: El proceso de ingeniería de atributos se vio afectado por varios datos erróneos y problemas de parseo (ej. ‘\n’ en cualquier parte de los datos)
- Experiencia: El equipo no cuenta con experiencia previa con el estilo y herramientas del proyecto
- Capacidad de cómputo y velocidad de internet: Limita la cantidad de datos y consultas a realizar

Arquitectura

Fig. 7

Resultados

Gasto total de los meses analizados (7 meses):

- **US$ 17,935,363,328**
- **Promedio (US$ 2,562,194,761)**

Gasto promedio de una compra por mes:

- **US$ 7,380**

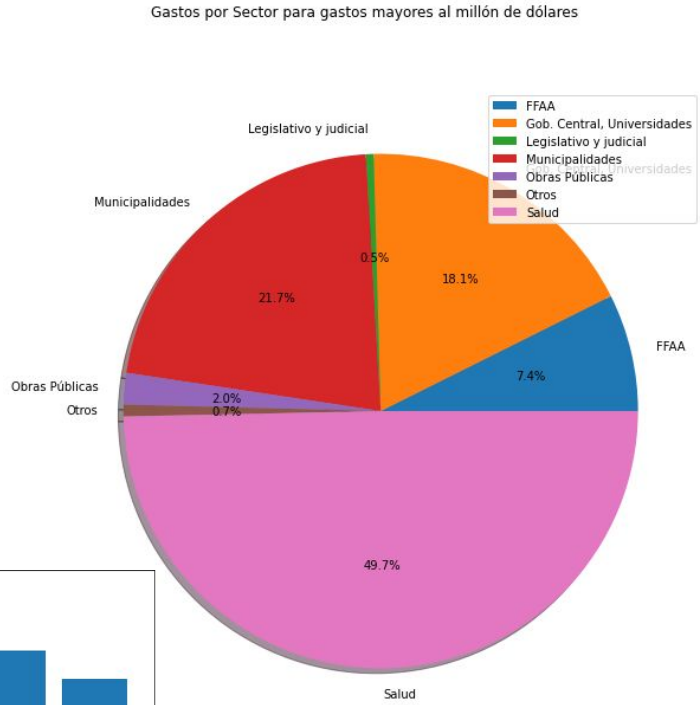

Fig. 2

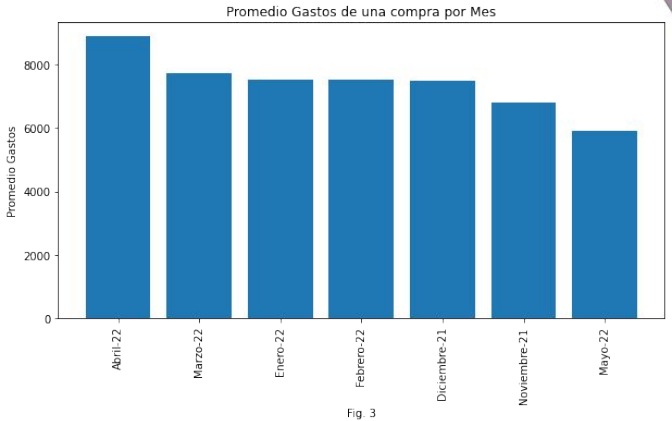

Fig. 3

Resultados

Mayor gasto en una región:

US$ 8,779,409,333

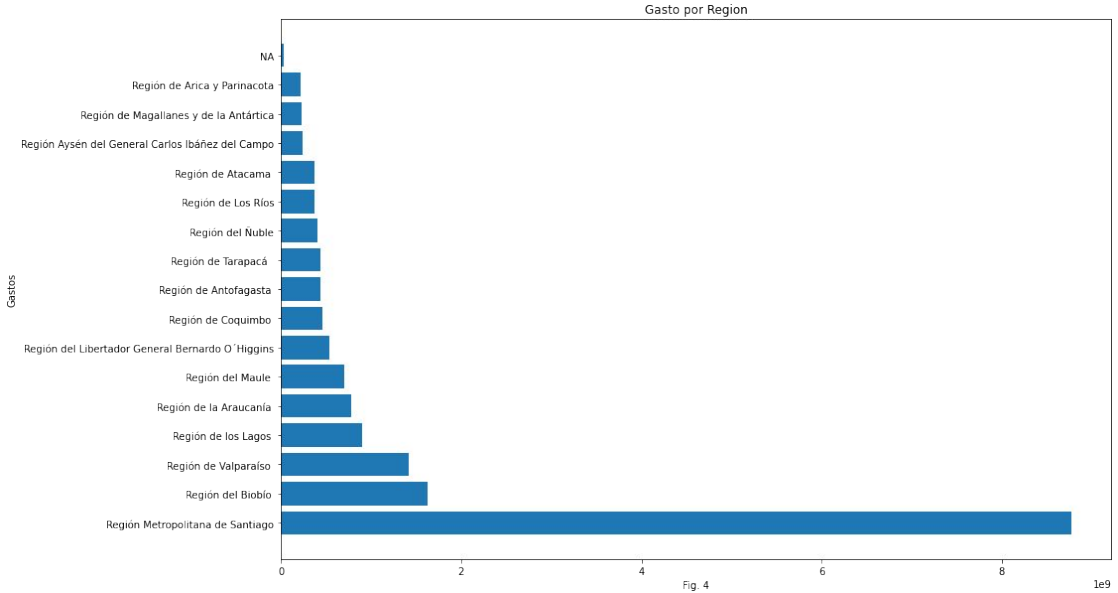

Fig. 4

La región metropolitana de Santiago es la que tuvo mayores gastos y esto es acorde a que allí se encuentra la capital del país. La desproporción es notoria.

Conclusiones

1. Se logró utilizar herramientas para análisis de grandes datos e implementar la arquitectura planteada
2. Se creo un dataset a partir de datos de varios meses
3. Realizando ingeniería de atributos se logró llegar a un dataset donde se pueden hacer consultas útiles. Se determinaron y eliminaron los atributos que no eran relevantes para las preguntas
4. Se analizaron los datos utilizando spark y se extendió el análisis de algunas respuestas
5. El análisis realizado muestra que Chile invierte desproporcionadamente entre sectores, siendo Salud el claro ganador
6. Se respondieron las consultas planteadas proporcionando gráficas y resúmenes usando spark y pandas
7. Se realizaron consultas a través del motor de búsqueda Solr.

Recomendaciones

1. Investigar otras soluciones/algoritmos si se quiere agrupar todos los meses en formatos csv del estado chileno en tiempos menores a 30 minutos.
2. Comparar con años anteriores si efectivamente la pandemia mundial de Covid-19 tuvo un efecto significativo en que las mayores inversiones de Chile sean en el sector de salud, investigando la respuesta a nuestra hipótesis en el tema.
3. Investigar/implementar otras herramientas para visualización de datos
4. Analizar y comparar los datos obtenidos con datos abiertos de otros países (ej. Uruguay)

Bibliografía

Bonillo, P. (2018). Propuesta de una Arquitectura de Gestión de Grandes Volúmenes de Datos para la Analítica en Tiempo Real bajo Software Libre. Link: https://www.linkedin.com/pulse/propuesta-de-una-arquitectura-gesti%C3%B3n-grandes-datos-la-bonillo-ramos/

Documentación de Apache Spark, Link: https://spark.apache.org/docs/latest/
Documentación de Papermill, Link: https://papermill.readthedocs.io/en/latest/

Documentación de Pandas, Link: https://pandas.pydata.org/