

Herramientas de software para Big Data

(Ingeniería en Sistemas - M7A)

Obligatorio 2022:

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

239850 – Iván Monjardin

219401 – Francisco Rossi

Índice

Introducción	2
Planteamiento del Problema	3
Objetivo General: Desarrollar un sistema de análisis de datos sobre las compras estatales (caso Chile) bajo Big Data.	3
Objetivos Específicos	3
Alcance	4
Propuesta de Valor	4
Metodología	5
Plan de Trabajo	5
Ingeniería de atributos	5
Análisis cualitativo	6
Análisis de nulidad y componentes estadísticos	9
Análisis de correlación	10
Arquitectura de la Solución	11
Conclusiones	13
Restricciones	24
Data set depurado	24
Archivos entregados	24
Bibliografía	25

Introducción

El siguiente documento representa el trabajo obligatorio realizado para la materia Herramientas de software para Big Data, donde se muestra un análisis de las operaciones que se realizaron.

Se discutió cual data set se utilizaría para el obligatorio, investigando diferentes fuentes y categorías. Se decidió seleccionar el dataset correspondiente a las compras del estado de Chile ya que, al querer hacer el mismo análisis sobre las compras del estado uruguayo, no contaba con el millón de registro, como el obligatorio solicitaba.

<http://datosabiertos.chilecompra.cl/Home/DescargaHistorico>

Para llegar al millón de registros, utilizamos todos los datos desde noviembre de 2021 y los del 2022 hasta el momento (mayo 2022).

Los campos más interesantes para analizar en un principio son, los ítems con sus respectivos valores de compras, sus proveedores y compradores. También las categorías de los productos que se destacan y sus tendencias.

La siguiente tabla resumen los atributos más relevantes de los 72 disponibles:

ID	Identificador único de la compra
FechaCreacion	Fecha en la que fue creado el pedido de compra
FechaAceptacion	Fecha de cuando fue aprobada la compra
MontoTotalOC_PesosChilenos	Monto de la orden de compra en pesos chilenos
sector	A qué área está destinada la compra
NombreroductoGenerico	Nombre del producto
OrganismoPublico	Organismo que realiza la compra
Categoria	Categoría del producto/servicio
TotalNetoOC	Precio total en moneda original luego de impuestos
CodigoOrganismoPublico	Identificador del organismo público
RegionUnidadCompra	Región de Chile donde se realiza compra
PaisProveedor	País donde se compra el producto
CiudadUnidadCompra	Ciudad destino de Orden de compra
cantidad	Cantidad de ítems

precioNeto	Precio en moneda original por cantidad de unidades
------------	--

Planteamiento del Problema

Utilizando los datos obtenidos se buscará mostrar de manera más intuitiva y transparente 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 puede ser completa, realizar consultas sobre ella no siempre es fácil.

Se intentará representar los datos del dataset utilizando gráficas y otras herramientas para mostrar de manera más intuitiva y resumida los datos.

Los resultados del análisis permitirán una mayor transparencia hacia el público en general e interesados.

Preguntas de investigación a resolver:

- ¿Cuál es el gasto total del estado chileno en los meses analizados?
- ¿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?
- ¿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?
- ¿En qué ciudades/regiones se realizaron los mayores gastos?
- ¿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?

Objetivo General: Desarrollar un sistema de análisis de datos sobre las compras estatales (caso Chile) bajo Big Data.

Utilizar gráficas y otros elementos visuales para mostrar resúmenes sobre las compras del estado chileno. Además se realizará un análisis sobre los gastos públicos, clasificándolos por diferentes categorías, por ejemplo gastos mayores a 1 millón de dólares si es que existen suficientes datos.

Objetivos Específicos

Se plantean los siguientes objetivos específicos:

- Realizar ingeniería de atributos, analizando y determinando atributos que pueden tener mayor valor para consultas que se quieran realizar.

- Definir si conviene más utilizar los datos en formato CSV o [OCDS](#) (ambos formatos están disponibles)
- Determinar e implementar una solución para almacenamiento de los datos adecuada para la cantidad y formato de los datos, además de los análisis que se quieran realizar.
- Desarrollar una arquitectura de Big Data que permita el análisis de compras del estado chileno
- Determinar gráficas o consultas que se quieren responder
 - Ej. Mayores gastos por región de Chile
- Utilizar herramientas para análisis de grandes datos, por ejemplo Spark + Jupyter Notebooks, para resolver las consultas del punto anterior.
- Responder a las preguntas de investigación.
- Utilizar un motor de búsqueda. Herramienta Solr

Alcance

El presente trabajo representa un ensayo teórico, basado en datos reales.

Su objetivo es validar los conceptos, metodologías y herramientas vistas a lo largo del curso, y por lo tanto, se limitará a:

- Realizar un informe con gráficas y resúmenes interesantes sobre los datos que respondan a las preguntas de investigación previamente definidas.
- Permitir realizar análisis sobre los datos con el fin de responder las preguntas de investigación.
- Responder cuáles fueron los ítems más comprados, los promedios de gastos y mayores gastos en general.
- Análisis únicamente de un subconjunto de los datos obtenidos en <http://datosabiertos.chilecompra.cl/Home/DescargaHistorico> en la sección “Reporte de Ordenes de Compra”

No se incluye en este enfoque:

- Evaluación de las mejores herramientas a utilizar para este tipo de análisis.
- Verificación o comprobación de los datos obtenidos con la realidad del estado chileno.
- Mejoramiento o enriquecimiento de los datos en caso de ser necesario.
- Plan de inversión a futuro para el estado chileno.
- Conversión exacta con respecto a las fechas, entre el peso chileno y dólares
- Análisis de las Licitaciones - Sólo se analizarán Órdenes de Compra
- Proponer acciones y políticas en base a los resultados obtenidos
- No se realizará un modelo de Machine Learning ya que no se tiene un único objetivo final para tratar de responder

Propuesta de Valor

Los gobiernos (tanto el de Chile, Uruguay entre otros) hacen públicos los datos de sus compras, sin embargo no siempre los aprovechamos al máximo. Algunos datos no son accesibles sin un análisis más profundo y un expertise técnico.

Nuestro trabajo propone realizar ese tipo de análisis para resolver algunas de esas consultas que consideramos relevantes para el público, facilitando la información ya propuesta por el estado, de manera que se pueda obtener un acercamiento más ameno a los mismos.

El estado tiene la responsabilidad de realizar un gasto del dinero ordenado y responsable con respecto a inversiones y compras, más aún en épocas de crisis como el de una pandemia mundial, queremos ver como se refleja esto en el período analizado.

También, nuestro trabajo tiene un enfoque social, la publicación de resúmenes de datos del estado permite a los posibles lectores tener un mejor entendimiento de la situación del país el cual puede ser diferente a lo que cada individuo podría estimar.

Si bien este proyecto se enfoca en un periodo específico de tiempo, este trabajo puede servir como base para futuros análisis de los datos que se publiquen a futuro y de todos los datos que se han publicado hasta ahora. Permitiendo realizar un análisis con datos más completos y actualizados.

Metodología

El presente trabajo se realizará utilizando metodologías de desarrollo ágil, siguiendo los principios ágiles, como priorizar la satisfacción del cliente (propuesta de valor y cumplimiento del obligatorio), aceptar cambios de requerimientos durante el desarrollo, software funcionando como medida de avance y otros.

Si bien no utilizaremos Scrum, Pedro Bonillo efectuará de Product Owner y Scrum Master con reuniones weekly en vez de dailys.

Los entregables del trabajo son:

- Informe de trabajo realizado con documentación de ingeniería de atributos realizados
- Presentación Power Point para la defensa
- Jupyter Notebook y cualquier otro artefacto de código o similar que desarrollemos)
- Data set depurado en formato csv

Plan de Trabajo

Al momento de realizar el proyecto se realizan las siguientes actividades:

- Depuración de datos (ingeniería de atributos) y selección de atributos relevantes.
- Definir análisis concretos que se desean realizar sobre los datos
- Definir una arquitectura de Big Data que permita dichos análisis
- Implementar la arquitectura
- Se analizan los resultados y se realiza análisis de compras del estado según diferentes categorías, con el objetivo de responder las preguntas de investigación.
- Realizar un resumen de los análisis realizados con las fechas propuestas.

Ingeniería de atributos

A continuación se realizan los siguientes análisis sobre los atributos del dataset:

- Análisis cualitativo de las variables a utilizar, en base a la importancia y utilizada para responder las consultas que se quieren realizar.
- Análisis de campos nulos y componentes estadísticos.
- Análisis de correlación entre variables para evitar errores de ponderación en el modelo.

Análisis cualitativo

Atributo	Descripción
ID	Index - No aplica para el análisis
Código	No aplica - Asumimos que no tiene poder explicativo
Link	No aplica - Asumimos que no tiene poder explicativo
Nombre	No aplica - Se utilizarán otros atributos para determinar el tipo de compra
Descripcion/Obervaciones	No aplica - Asumimos que no tiene poder explicativo
Tipo	No aplica - Asumimos que no tiene poder explicativo
ProcedenciaOC	No aplica - Asumimos que no tiene poder explicativo
EsTratoDirecto	No aplica para consultas que queremos realizar
EsCompraAgil	No aplica para consultas que queremos realizar
CodigoTipo	No aplica - Asumimos que no tiene poder explicativo
CodigoAbreviadoTipoOC	No aplica para consultas que queremos realizar
DescripcionTipoOC	No aplica para consultas que queremos realizar
codigoEstado	No aplica - Asumimos que no tiene poder explicativo
Estado	Se utilizara para saber si la solicitud está aceptada o no
codigoEstadoProveedor	No aplica - Asumimos que no tiene poder explicativo
EstadoProveedor	No aplica para consultas que queremos realizar
FechaCreacion	Se utilizará para determinar tiempo desde que se ordenó hasta que se concretó
FechaEnvio	No aplica para consultas que queremos realizar
FechaSolicitudCancelacion	No aplica para consultas que queremos realizar
fechaUltimaModificacion	No aplica para consultas que queremos realizar
FechaAceptacion	Se utilizará para determinar tiempo desde que se ordenó hasta que se concretó
FechaCancelacion	No aplica para consultas que queremos realizar
tieneItems	No aplica - Asumimos que no tiene poder explicativo

PromedioCalificacion	No aplica para consultas que queremos realizar
CantidadEvaluacion	No aplica para consultas que queremos realizar
MontoTotalOC	No aplica para consultas que queremos realizar
TipoMonedaOC	No aplica para consultas que queremos realizar
MontoTotalOC_PesosChile nos	Se utilizará para analizar las compras de mayor gasto y los promedios de las mismas
Impuestos	No aplica para consultas que queremos realizar
TipoImpuesto	No aplica para consultas que queremos realizar
Descuentos	No aplica para consultas que queremos realizar
Cargos	No aplica para consultas que queremos realizar
TotalNetoOC	No aplica - Asumimos que no tiene poder explicativo
CodigoUnidadCompra	No aplica - Asumimos que no tiene poder explicativo
RutUnidadCompra	No aplica - Asumimos que no tiene poder explicativo
UnidadCompra	No aplica - Asumimos que no tiene poder explicativo
CodigoOrganismoPublico	No aplica para consultas que queremos realizar
OrganismoPublico	Se utilizará para clasificar compras según organismo
sector	Se utilizará para clasificar las compras según sectores
ActividadComprador	No aplica para consultas que queremos realizar
CiudadUnidadCompra	Se utilizará para identificar las ciudades que realizan mayor cantidad de compras
RegionUnidadCompra	Se utilizará para identificar las región que realizan mayor cantidad de compras
PaisUnidadCompra	No aplica - Asumimos que no tiene poder explicativo
CodigoSucursal	No aplica - Asumimos que no tiene poder explicativo
RutSucursal	No aplica - Asumimos que no tiene poder explicativo
Sucursal	No aplica para consultas que queremos realizar, se utilizará atributos como Organismo Público para diferenciar comparadores
CodigoProveedor	No aplica - Asumimos que no tiene poder explicativo
NombreProveedor	No aplica para consultas que queremos realizar
ActividadProveedor	No aplica para consultas que queremos realizar
ComunaProveedor	No aplica para consultas que queremos realizar
RegionProveedor	No aplica para consultas que queremos realizar
PaisProveedor	Se utilizará para analizar de donde provienen la mayor cantidad de compras
Financiamiento	No aplica para consultas que queremos realizar

PorcentajeIva	No aplica para consultas que queremos realizar
Pais	No aplica - Asumimos que no aporta poder explicativo
TipoDespacho	No aplica - Asumimos que no tiene poder explicativo
FormaPago	No aplica para consultas que queremos realizar
CodigoLicitacion	No aplica - Asumimos que no tiene poder explicativo
Codigo_ConvenioMarco	No aplica - Asumimos que no tiene poder explicativo
IDItem	Se utilizará para determinar cuál fue el item más comprado
codigoCategoria	No aplica - Asumimos que no tiene poder explicativo
Categoria	No aplica - Es lo mismo que "RubroN1", "RubroN2", "RubroN3" concatenado
codigoProductoONU	No aplica - Asumimos que no tiene poder explicativo
NombreroductoGenerico	No aplica para consultas que queremos realizar
RubroN1	Se utilizará para determinar en que invierten más las diferentes áreas
RubroN2	Se utilizará para determinar las secciones de compras
RubroN3	Se utilizará para determinar las secciones de compras
EspecificacionComprador	No aplica para consultas que queremos realizar
EspecificacionProveedor	No aplica para consultas que queremos realizar
cantidad	Se utilizará para ver qué compras se repitieron más
UnidadMedida	Se utilizará para ver que se compró más según la unidad.
monedaltem	No aplica - Asumimos que no tiene poder explicativo
precioNeto	No aplica para consultas que queremos realizar
totalCargos	No aplica para consultas que queremos realizar
totalDescuentos	No aplica para consultas que queremos realizar
totalImpuestos	No aplica para consultas que queremos realizar
totalLineaNeto	No aplica para consultas que queremos realizar
Forma de Pago	No aplica para consultas que queremos realizar

Análisis de nulidad y componentes estadísticos

Se presentan los principales hallazgos del análisis realizado:

- Depuración de registros nulos
Se eliminaron los valores nulos encontrados en el dataset para la columna de FechaAceptación. Se encontraron nulos en otros campos, como por ejemplo PaísProveedor, pero se asumió que se trataba de Chile y no se eliminaron esos datos.
- Se eliminaron algunos registros que contenían caracteres como “;” y comillas entre sus valores, lo que dificulta la importación de datos
- Para corregir los registros que, durante la lectura del csv, habían sido separado en varias líneas, se verificó la cantidad de elementos por líneas y se corrigieron o eliminaron las que no contenían 78 valores (cantidad de atributos del data set)
- Se eliminaron los registros que no contenían valores válidos para el atributo “Estado”, de esta manera si hubo algún error en la importación, la mayoría de registros erróneos se eliminan son eliminados.
- Para el análisis en Solr se eliminaron todos los datos nulos y se creó un función para convertir los datos del csv depurado desde spark a un csv que puede ser importado por Solr. El principal problema eran los caracteres especiales ya que spark y Solr tienen distintas convenciones para ignorar dichos caracteres lo cual provocaba problemas. Por lo tanto se eliminaron todos los caracteres como las comillas y las ‘ /” ‘

Análisis de correlación

Tal como se comentó previamente se presenta un análisis de correlación para las variables seleccionadas.

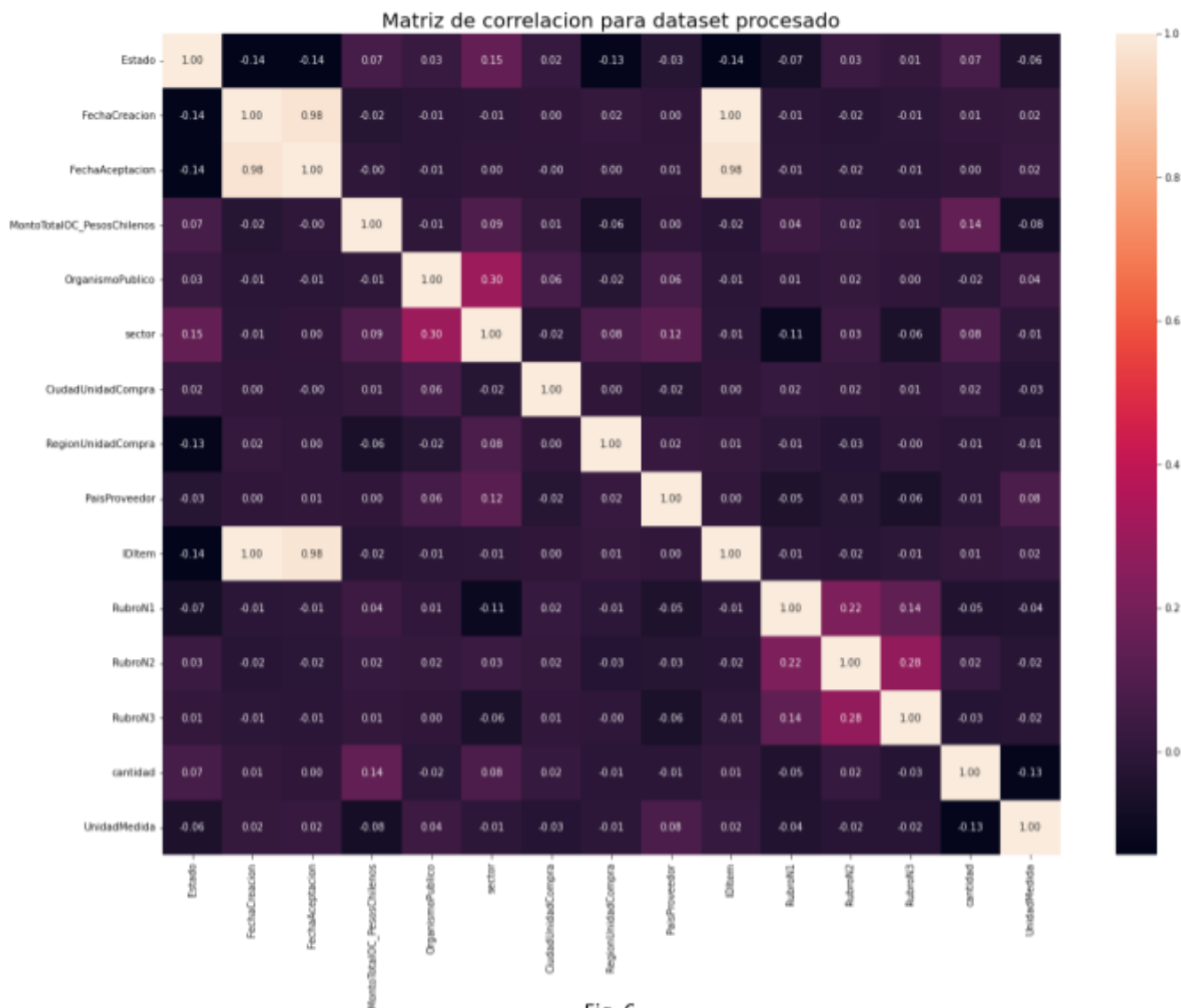


Fig. 6

A partir de la matriz, se puede observar que las únicas correlación significativa entre las variables es entre IDItem y FechaCreacion de la orden de compra, y entre ambos campos de fechas, pero decidimos continuar a realizar las consultas pertinentes con todos los campos disponibles ya que no aporta valor quitar estos elementos.

Arquitectura de la Solución

Para poder explicar mejor nuestra solución, proponemos la siguiente imagen que modela nuestra arquitectura.

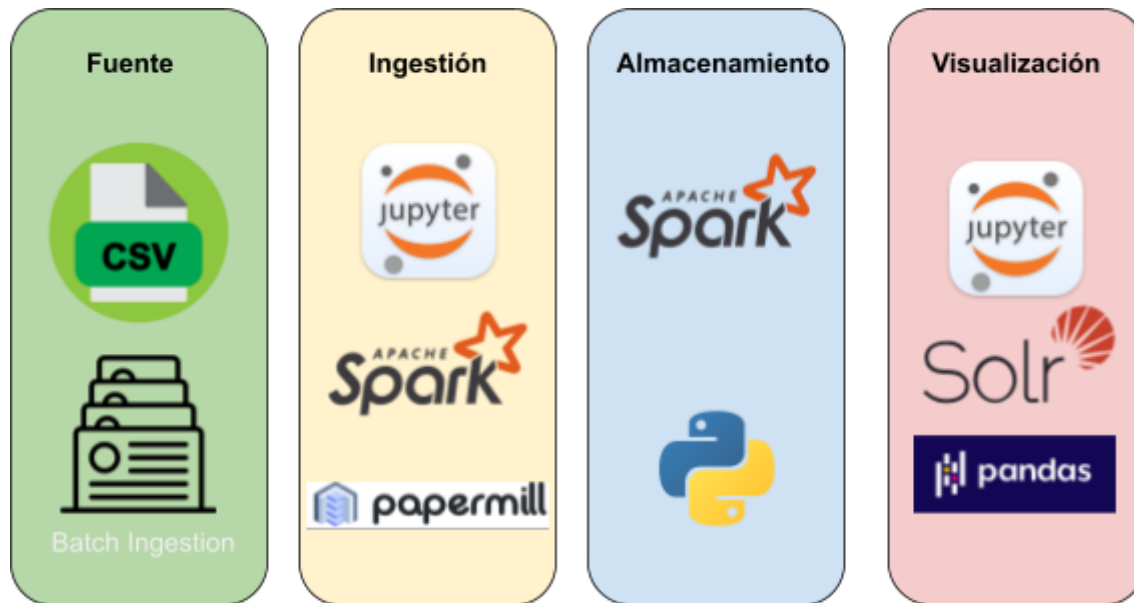


Fig. 7

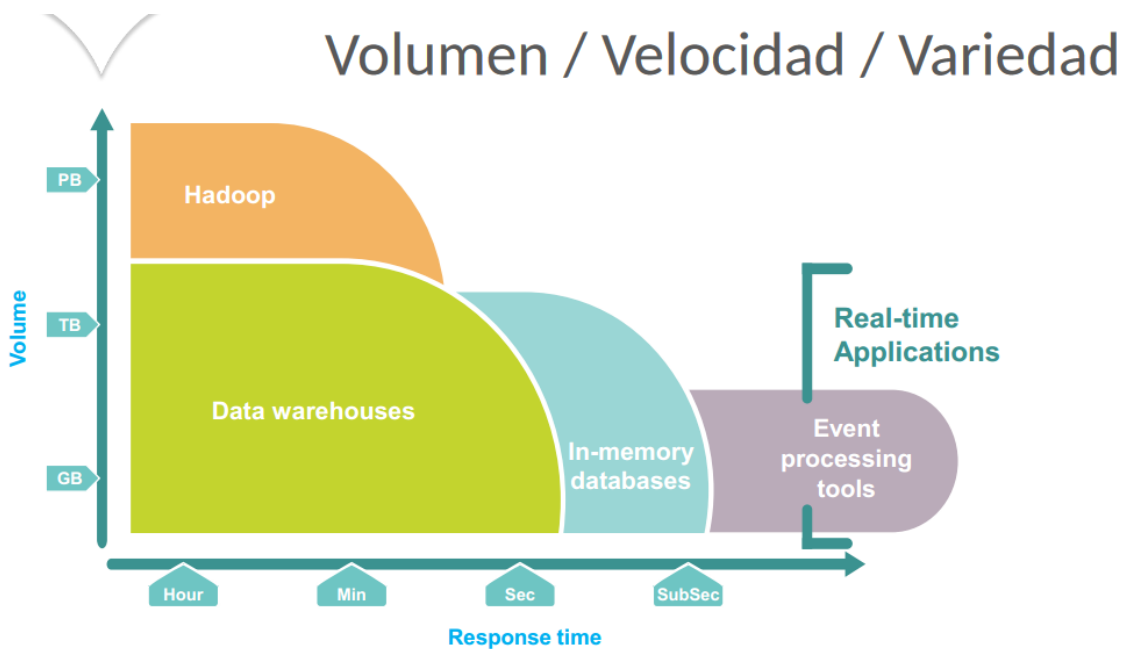
Pasos del proceso:

- 1- Los datos se toman desde la página oficial de compras del estado chileno, descargando mes a mes los datasets correspondientes al año 2021 y 2022 (hasta marzo, inclusive) en formato csv. Se procesan cada uno de los archivos, corrigiendo errores de formato y uniéndolos en un solo archivo con la herramienta papermill (ejecutando una notebook por mes) . Se utilizan estas herramientas mencionadas, por la capacidad de manejar grandes volúmenes de datos, ya que el trabajo se basa en la idea de poder manejar al menos 1 millón de datos de forma performante.
Tomamos 7 meses para este trabajo que son Mayo, Abril, Marzo, Febrero, Enero de 2022 y Diciembre, Noviembre de 2021, alcanzando 2.5 millones de registros.
- 2- Utilizando la herramienta spark se realiza la ingeniería de atributos, seleccionando 15 de estos considerados significativos.
Esta herramienta nos permitió realizar un proceso de depuración del dataset robusto, ya que contaba con errores de formato como también campos nulos. Esta sección del trabajo fue sustancialmente más difícil de lo que se pensó en un principio, ya que el formato del csv no era correctamente convertido a un dataframe. Esto requerirá definir funciones en python propias que pudieras solucionarlo y adecuarse al problema.
Adicionalmente spark es una herramienta abierta y robusta, que no requiere inversión en licencia.
- 3- Se utiliza la librería de Pandas, únicamente para realizar gráficas, una matriz de correlación de datos y también para realizar un reporte sobre el dataset que puede ser encontrado como un archivo anexo bajo el nombre "ComprasChile2021-2022.html". Estos ploteos son

frecuentemente utilizados en la industria, por lo que consideramos importantes agregarlos al informe y a la notebook.

- 4- Las consultas que se realizaron con el objetivo de responder las preguntas planteadas en un principio. Se hicieron con la herramienta de pyspark específica para consultas sql *"pyspark.sql.function"*. Esto se debe a que es simple de usar, con una amplia documentación, pero también la notamos performante.
- 5- Incluimos un motor de búsqueda como es Solr para poder visualizar rápidamente y realizar búsquedas sobre los datos ya depurados.

Tomando en cuenta la siguiente imagen podemos clasificar nuestros datos:



Los datos que importamos y analizamos usando batch tienen un peso aproximado de medio GB y es analizado en minutos, por lo tanto entra en la categoría de Data warehouses, sin embargo nosotros los analizamos con Spark y Jupyter Notebooks. Además, tomando en cuenta que subimos los datos a Solr y son analizados en menos de un segundo, entraría en la zona de Event processing tools, respondiendo a consultas en tiempo real o cercano a tiempo real..

Conclusiones

Las conclusiones obtenidas son:

Realizar ingeniería de atributos, analizando y determinando atributos que pueden tener mayor valor para consultas que se quieran realizar.

Luego de un análisis cualitativo y cuantitativo de los atributos se decidió seleccionar los siguientes atributos:

- Estado
- FechaCreacion
- FechaAceptacion
- MontoTotalOC_PesosChilenos
- OrganismoPublico
- sector
- CiudadUnidadCompra
- RegionUnidadCompra
- PaisProveedor
- IDItem
- RubroN1
- RubroN2
- RubroN3
- cantidad
- UnidadMedida

El Estado sirve para saber si la compra fue aceptada, cancelada, incompleta, si ya fue enviada o si ya fue recibida.

Las fechas fueron seleccionadas debido a que se quiere responder preguntas en base a esta información, como por ejemplo gastos del estado en promedio por mes.

MontoTotalOC_PesosChilenos sirvió para analizar varias de las preguntas claves, por lo que se podría tratar de uno de los campos más importante del dataset. Una nota a mencionar es que las funciones de pyspark se encuentran en formato americano, por lo luego de investigar, nos dimos cuenta que al castear a tipo “double”, contenía errores ya que el carácter coma, no lo reconocía correctamente, generando más de 10000 nulos por documento. Al darnos cuenta de eso pudimos solucionarlo y contar con todos los valores disponibles.

Sector y RegiónUnidadCompra nos ayudaron a entender de forma más clara la distribución de los gastos que se realizó por el estado chileno, pudiendo sacar algunas hipótesis propias.

Se tuvo que realizar una depuración de formato bastante intensa sobre los datos originales csv antes de pasarlos a dataframe. Esto supuso un intenso trabajo de investigación ya que no se podía arreglar simplemente con un regex como es habitual, dado estos errores de formato específicos. Este proceso nos tomó varios días de trabajo, algo que no habíamos considerado a la hora de planificar la ingeniería de atributos.

Adicionalmente fue necesario depurar menos del 6% de los datos disponibles por casos de nulidad o valores inconsistentes. Por ejemplo se eliminaron las líneas que contenían valores inválidos en la columna de Estado, FechaAceptación nulas.

Finalmente, se determinó que con una matriz de correlación previamente mostrada en el informe que las variables seleccionadas no cuentan con correlación significativa, por lo que se puede proseguir con el análisis.

Definir si conviene más utilizar los datos en formato CSV o OCDS (ambos formato están disponibles)

En un principio analizamos la posibilidad de utilizar OCDS como formato plano de los datos antes de procesarlos y convertirlo a data frame. Pero consultando con referentes y documentación, se observó que csv es un estándar comúnmente utilizado en esta área. Esto implica que existen funciones pre definidas que están habilitadas para hacer conversión directa a data frame y la documentación es ampliamente mayor a la del formato OCDS.

Aún así, como ya se mencionó, ocurrieron varios problemas con el formato csv, pero creemos que esto puede ocurrir en la industria seguido, el tener que corregir los datos intensamente antes de poder trabajar con ellos y sacar resultados.

Desarrollar una arquitectura de Big Data que permita el análisis de compras del estado

Se creó una arquitectura de Big Data mencionada previamente en el documento, donde se utilizaron distintas herramientas para poder completar la investigación.

Entre ellas se encuentran:

- Archivos CSV
- Apache Spark
- Papermill
- Jupyter Notebooks
- Python
- Pandas
- Solr

Utilizar herramientas para análisis de grandes datos, por ejemplo Spark + Jupyter Notebooks, para resolver las consultas del punto anterior.

Se utilizaron los servidores de la Universidad ORT, para no tener que procesar todos los datos de forma local, siendo estos 2.5 millones de tuplas. Esto ayudó significativamente los tiempos de ejecución.

Sin embargo cabe aclarar que el procesamiento de los archivos de datos para generar el batch tarda aproximadamente 25 minutos. Es un tiempo a tener en cuenta, pero esto se realiza una única vez para procesar todos los 7 meses analizados, y se quisiera agregar un mes más, se estima que tardaría alrededor de 3 minutos.

Entendemos que al realizarse una única vez por mes, son tiempos aceptables para procesar millones de datos.

Responder a las preguntas de investigación incluyendo gráficas en algunos casos

A continuación se responderán las preguntas que se plantearon antes de comenzar el informe, sobre preguntas que creemos son relevantes dado el dataset que se cuenta.

Los datos que corresponden a gastos, se mostrarán en dólares americanos para una mejor comprensión de los mismos.

- ¿Cuál es el gasto total del estado chileno en los meses analizados?

Total de Gastos del estado chileno fue de :

US\$ 17,935,363,328

- ¿Qué áreas está el Estado chileno invirtiendo más?

Por Sector:

Los mayores gastos que se vieron por sector fueron:

- Salud (US\$ 8,907,535,239)
- Municipalidades (US\$ 3,891,131,770)
- Gob. Central, Universidades (US\$ 3,240,210,412)

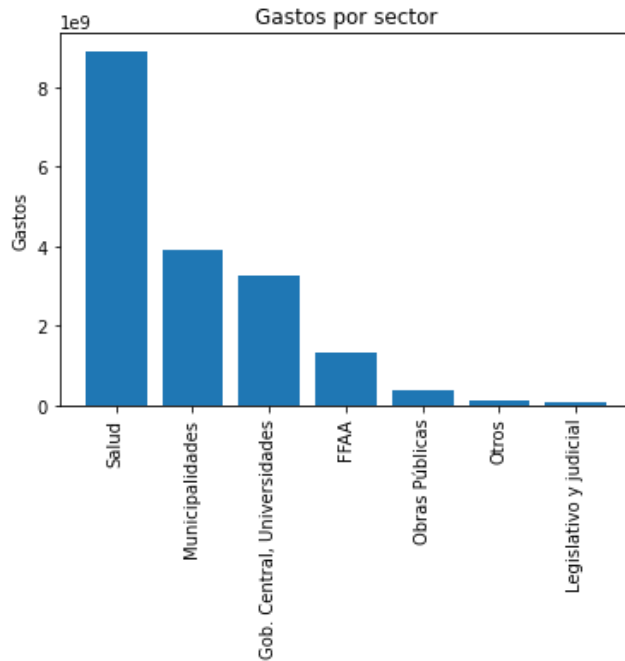


Fig. 1

Podemos sacar una hipótesis de que se gastó sustancialmente más en salud ya que se transcurrió una pandemia a nivel mundial. Habría que comparar e investigar con meses previos a la pandemia para confirmar este supuesto.

Por Rubro:

Los rubros generales (columna Rubro1) que mayores gastos realizaron fueron:

- Equipamiento y suministros médicos (US\$ 2,854,918,768)
 - Equipamiento para laboratorios (US\$ 2,291,445,398)
 - Salud, servicios sanitarios y alimentación (US\$ 1,571,860,421)
- ¿Cuántos gastos mayores al millón de dólares fueron efectuados entre noviembre de 2021 y mayo de 2022?

Hay 1002 gastos mayores al millón de dólares. Siendo los mayores gastos por Rubros específicos (columna Rubro3):

- US\$ 180,793,734 → Plantas y árboles ornamentales
- US\$ 160,531,140 → Accesorios para herramientas
- US\$ 137,111,023 → Riego

Para observar estos resultados de gastos mayores al millón de dólares por sector, incluimos esta gráfica que distingue claramente la distribución de los gastos por sector.:

Gastos por Sector para gastos mayores al millón de dólares

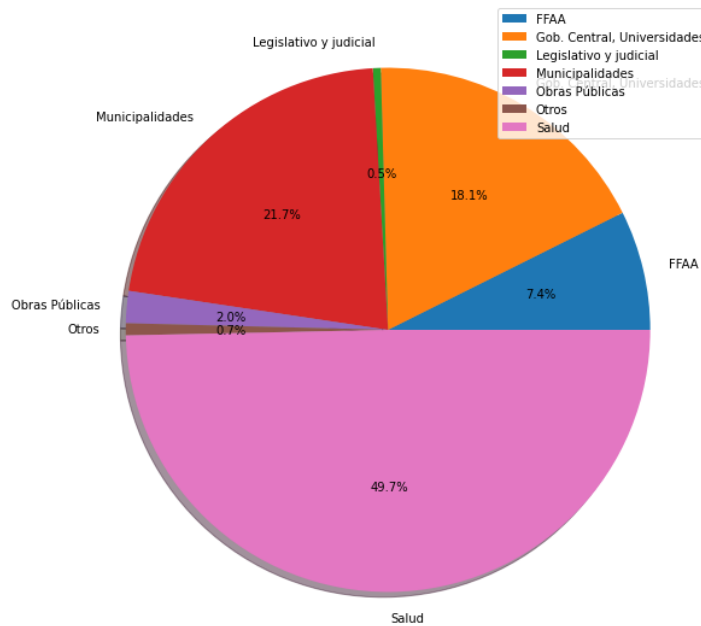


Fig. 2

Aunque los mayores gastos por Rubros fueron realizados en Plantas y árboles ornamentales, el sector que tuvo más gastos acumulados fue el de la salud

- ¿Cuál fue el gasto promedio en general? ¿Cuál fue el gasto promedio de una compra por mes del estado?

El gasto promedio en general de todas las compras fue de US\$ 7,380

Por mes el gasto promedio de una compra es el siguiente:

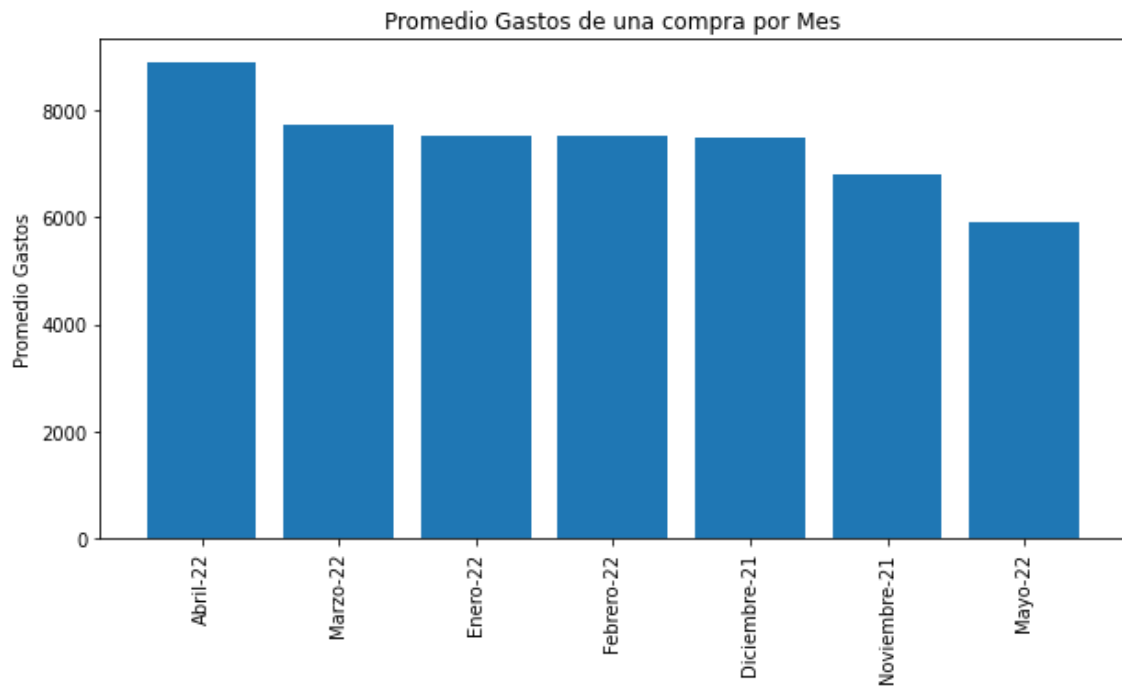


Fig. 3

A pesar de que pudimos observar en la anterior gráfica que hay gastos millonarios, podemos observar que el promedio es relativamente bajo, alrededor de US\$7500 por mes, se concluye que hay muchas compras relativamente pequeñas.

- ¿Cuál fue la compra más repetida del estado?

Analizamos esta pregunta pero el máximo que se repitió una compra fue 3, por lo tanto con los datos que tenemos no consideramos muy relevante esta pregunta.

- ¿En qué ciudades/regiones se realizaron los mayores gastos?

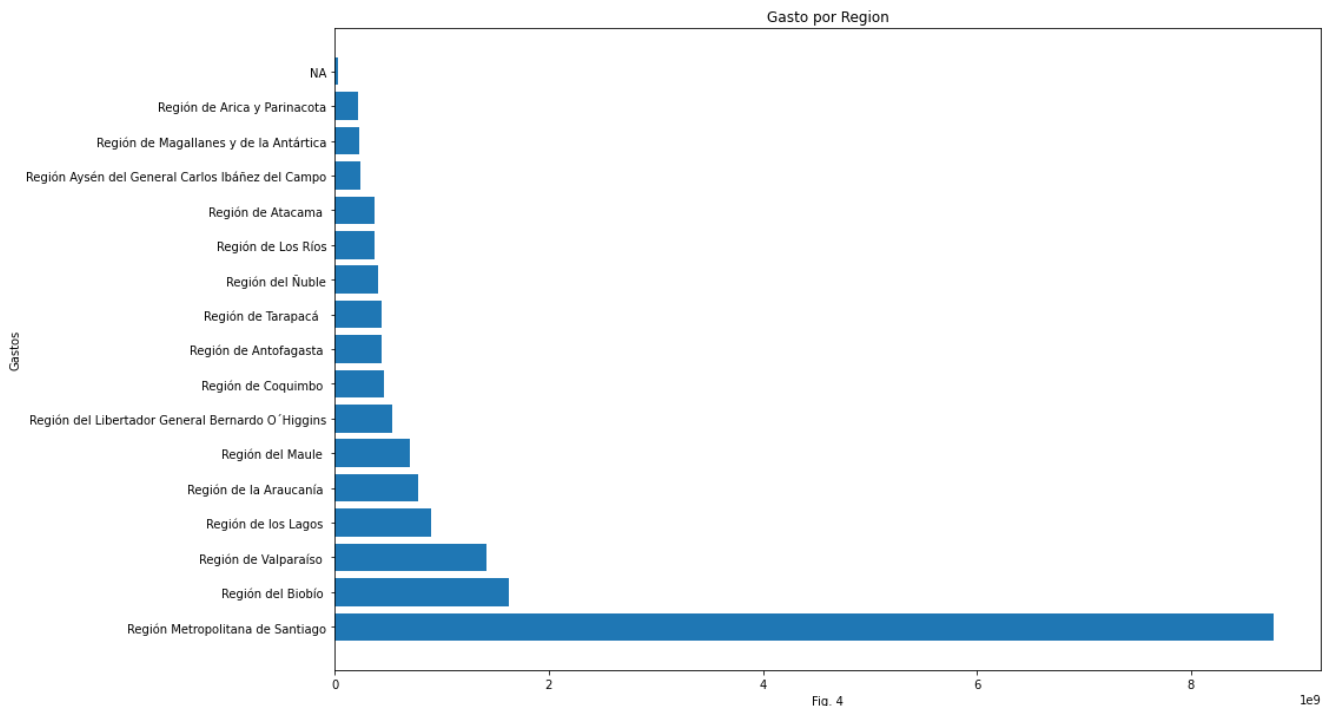
Por ciudad:

La ciudad que más se gastó fue en Santiago de Chile (US\$ 3,367,062,852).

Estos resultados son acordes a lo esperado, ya que se trata de la capital del País.

Por región:

En la región metropolitana de Santiago se realizaron los mayores gastos del estado chileno (US\$ 8,779,409,333).



Se investigó los gastos por regiones ya que se quiso comprobar si coincidía con otros gastos analizados por ciudades, donde se concluyó que la capital, Santiago de Chile es donde se realizaron mayores gastos en el período analizado. Se puede distinguir claramente en la gráfica que los datos de regiones coinciden con los datos obtenidos en los datos de gastos por ciudades.

Al realizar un análisis más profundo, podemos ver que se vuelve a repetir que salud es el sector con mayores gastos para la Región Metropolitana de Santiago de Chile.

Gastos por Sector para la Región Metropolitana de Santiago

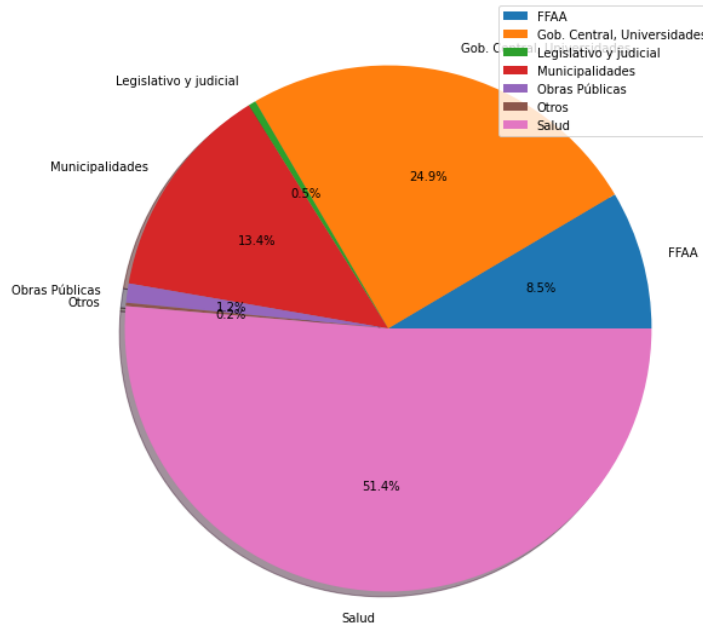


Fig. 5

- ¿Cuál es el tiempo promedio desde que se solicita la orden de compra, hasta que se autorice?

Desde la creación hasta la aceptación de la solicitud de compra en promedio demora 7 días.

- ¿Cómo se distribuyen las compras en función de organismos?

El organismo con mayores gastos fue:

- CENTRAL DE ABASTECIMIENTO DEL SISTEMA NACIONAL DE SERVICIO DE SALUD (US\$ 1,067,028,806)
- FONDO NACIONAL DE SALUD (US\$ 839,484,260)

Sin considerar los relacionados con la salud y el organismo con mayor gasto es JUNTA NACIONAL DE AUXILIO ESCOLAR Y BECA (US\$ 645,381,518)

En la siguiente imagen se puede observar la distribución de cantidad de compras por organismos para los primeros 20 Organismos Públicos del estado Chileno.

OrganismoPublico	Total Gastos
CENTRAL DE ABASTECIMIENTO DEL SISTEMA NACIONAL DE SERVICIO DE SALUD	US\$ 1,067,028,806
FONDO NACIONAL DE SALUD	US\$ 839,484,260
JUNTA NACIONAL DE AUXILIO ESCOLAR Y BECA	US\$ 645,381,518
SUBSECRETARIA DE SALUD PUBLICA	US\$ 609,783,173
I MUNICIPALIDAD DE VITACURA	US\$ 494,253,665
HOSPITAL GUILLERMO GRANT BENAVENTE DE CO	US\$ 293,177,793
DIRECCION DE ABASTECIMIENTO DE LA ARMADA	US\$ 287,608,120
DIRECCION DE LOGISTICA DE CARABINEROS	US\$ 276,195,196
SERVICIO DE SALUD METROPOLITANO SUR HOSP	US\$ 262,097,148
DIRECCION GENERAL DE GENDARMERIA DE CHIL	US\$ 260,976,617
COMANDO DE APOYO A LA FUERZA	US\$ 249,271,812
COMPLEJO ASISTENCIAL DR.VICTOR RIOS RUIZ	US\$ 244,236,127
SERVICIO DE SALUD DE TALCAHUANO HOSPITAL	US\$ 238,126,328
HOSPITAL PUERTO MONTT SERVICIO DE SALUD DEL RELONCAVI	US\$ 219,510,415
HOSPITAL CLINICO METROPOLITANO DE LA FLORIDA DOCTORA ELOISA DIAZ	US\$ 207,720,041
CORP NACIONAL FORESTAL	US\$ 202,578,399
MINISTERIO DE OBRAS PUBLICAS DIREC CION GRAL DE OO PP DCYF	US\$ 198,439,389
UNIVERSIDAD DE CHILE	US\$ 189,715,413
SERVICIO NACIONAL DE SALUD HOSPITAL CARLOS VAN BUREN	US\$ 186,914,204
JEFATURA DE AHORRO PARA LA VIVIENDA DEL EJERCITO.	US\$ 183,724,930

- ¿Qué proyecciones se pueden hacer a futuro sobre los datos analizados?

Como se pudo observar, el sector que más se invirtió fue el de salud. Creemos que esta tendencia va a continuar de esta forma, ya que la diferencia es sustancial, más de la mitad de los gastos fueron realizados en este sector. Aunque puede ser que la pandemia haya afectado significativamente estos datos, entendemos que la diferencia es tan grande que Salud es un prioridad para el estado chileno.

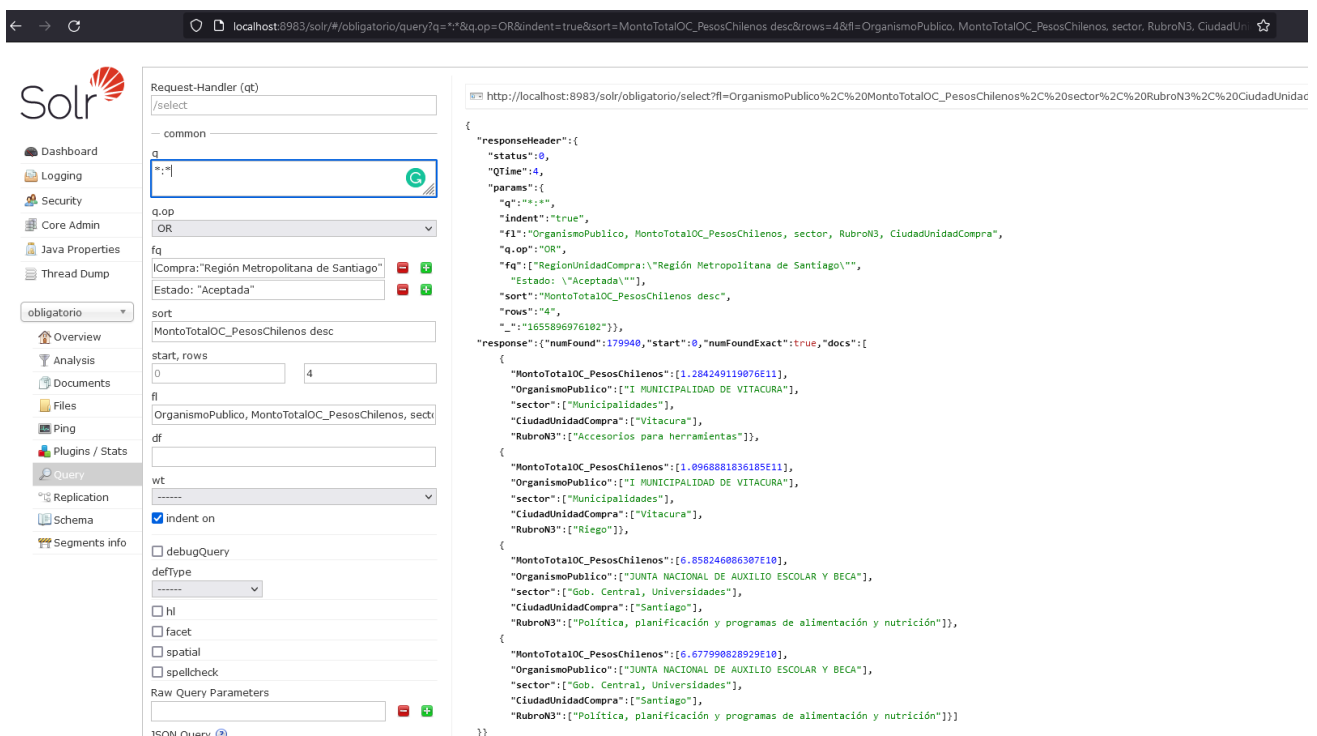
Utilizar un motor de búsqueda. Herramienta Solr

Nos propusimos poder utilizar un motor de búsqueda para los datos de las compras del Estado Chileno y para ello manejamos la herramienta vista en el curso Solr.

A pesar de que spark podía trabajar sin problemas con el data frame obtenido, Solr tuvo varios problemas a la hora de postear el archivo csv. Para resolver esto tuvimos que hacer una segunda depuración de los datos, reemplazando puntos y comas por comas, agregar un header que habíamos eliminado y eliminar alguna línea con caracteres especiales. También se decidió eliminar todos los nulos que quedaban en el dataset para prevenir errores inesperados.

Para trabajar con Solr, utilizamos una imagen de docker que nos facilitó muchas configuraciones previas.

Algunos ejemplos de queries que realizamos:



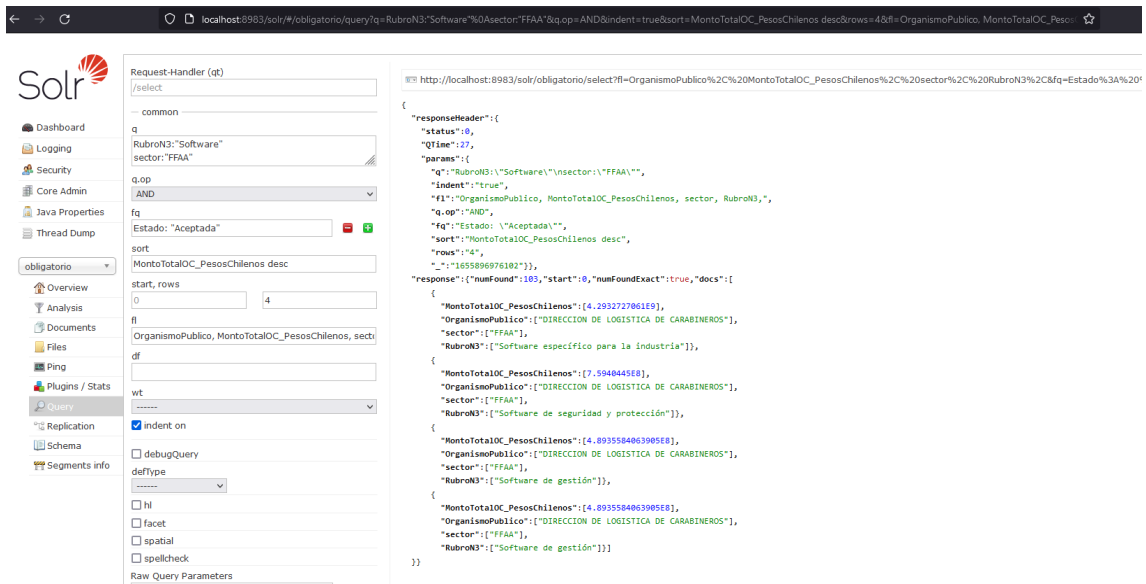
The screenshot shows the Solr Admin interface. On the left, the 'Query' tab is selected. The 'Request-Handler (qt)' is set to '/select'. The 'q' field contains the query: `*:*&q.op=OR&indent=true&sort=MontoTotalOC_PesosChilenos desc&rows=4&fl=OrganismoPublico, MontoTotalOC_PesosChilenos, sector, RubroN3, CiudadUnidadCompra&fq=RegionUnidadCompra:%22Región Metropolitana de Santiago%22&fq=Estado:%22Aceptada%22&qt=`. The 'q.op' is set to 'OR', 'fq' is 'RegionUnidadCompra: "Región Metropolitana de Santiago"', 'Estado: "Aceptada"', 'sort' is 'MontoTotalOC_PesosChilenos desc', 'start, rows' is '0 4', 'fl' is 'OrganismoPublico, MontoTotalOC_PesosChilenos, sector, RubroN3, CiudadUnidadCompra', 'df' is 'wt', 'wt' is 'indent on', 'debugQuery' is checked, 'defType' is 'Standard', 'hl' is checked, 'facet' is checked, 'spatial' is checked, 'spellcheck' is checked. The 'Raw Query Parameters' field is empty. On the right, the JSON response is displayed, showing a list of 4 records with fields like 'MontoTotalOC_PesosChilenos', 'OrganismoPublico', 'sector', 'CiudadUnidadCompra', and 'RubroN3'.

La query correspondiente es la siguiente:

http://localhost:8983/solr/#/obligatorio/query?q=.*&q.op=OR&indent=true&sort=MontoTotalOC_PesosChilenos%20desc&rows=4&fl=OrganismoPublico,%20MontoTotalOC_PesosChilenos,%20sector,%20RubroN3,%20CiudadUnidadCompra&fq=RegionUnidadCompra:%22Región Metropolitana de Santiago%22&fq=Estado:%22Aceptada%22&qt=

Aquí por ejemplo queríamos obtener las 4 mayores compras que fueron aceptadas para la región metropolitana de Santiago, solo mostrando algunos campos relevantes, como qué Organismo Público fue el responsable o en qué Rubro se gastó.

Otro ejemplo es:



The screenshot shows the Solr Admin interface for a cluster named 'obligatorio'. The 'Query' tab is active, displaying the following configuration:

- Request-Handler (qt):** /select
- q:** RubroN3:"Software" sector:"FFAA"
- q.op:** AND
- fq:** Estado:"Aceptada"
- sort:** MontoTotalOC_PesosChilenos desc
- start, rows:** 0, 4
- fl:** OrganismoPublico, MontoTotalOC_PesosChilenos, sector, RubroN3
- df:** (empty)
- wt:** (empty)
- indent on:** ☒
- debugQuery:** ☐
- deftype:** (empty)
- hl:** ☐
- facet:** ☐
- spatial:** ☐
- spellcheck:** ☐

The raw query parameters are: `http://localhost:8983/solr/#/obligatorio/select?q=RubroN3:"Software"&q.op=AND&indent=true&sort=MontoTotalOC_PesosChilenos desc&rows=4&fl=OrganismoPublico,MontoTotalOC_PesosChilenos,sector,RubroN3&fq=Estado:"Aceptada"&qt=/select`

The JSON response is as follows:

```
{
  "responseHeader": {
    "status": 0,
    "qtime": 27,
    "params": {
      "q": "RubroN3:\\\"Software\\\"\\nsector:\\\"FFAA\\\"\\n",
      "indent": "true",
      "fl": "OrganismoPublico, MontoTotalOC_PesosChilenos, sector, RubroN3",
      "q.op": "AND",
      "fq": "Estado: \\\"Aceptada\\\"",
      "sort": "MontoTotalOC_PesosChilenos desc",
      "rows": "4",
      "_": "1655896976102"
    }
  },
  "response": {
    "numFound": 103,
    "start": 0,
    "numFoundExact": true,
    "docs": [
      {
        "MontoTotalOC_PesosChilenos": 4.293272706119,
        "OrganismoPublico": "DIRECCION DE LOGISTICA DE CARABINEROS",
        "sector": "FFAA",
        "RubroN3": "Software espec\u00edfico para la industria"
      },
      {
        "MontoTotalOC_PesosChilenos": 7.594044568,
        "OrganismoPublico": "DIRECCION DE LOGISTICA DE CARABINEROS",
        "sector": "FFAA",
        "RubroN3": "Software de seguridad y protecci\u00f3n"
      },
      {
        "MontoTotalOC_PesosChilenos": 4.893558406390568,
        "OrganismoPublico": "DIRECCION DE LOGISTICA DE CARABINEROS",
        "sector": "FFAA",
        "RubroN3": "Software de gesti\u00f3n"
      },
      {
        "MontoTotalOC_PesosChilenos": 4.893558406390568,
        "OrganismoPublico": "DIRECCION DE LOGISTICA DE CARABINEROS",
        "sector": "FFAA",
        "RubroN3": "Software de gesti\u00f3n"
      }
    ]
  }
}
```

Query correspondiente:

http://localhost:8983/solr/#/obligatorio/query?q=RubroN3:%22Software%22%0Asector:%22FFAA%22&q.op=AND&indent=true&sort=MontoTotalOC_PesosChilenos%20desc&rows=4&fl=OrganismoPublico,%20MontoTotalOC_PesosChilenos,%20sector,%20RubroN3,&fq=Estado:%20%22Aceptada%22&qt=

Donde quisimos exigir un poco m\u00e1s al motor de b\u00fasqueda y buscar por palabras que un campo pod\u00eda contener. En esta b\u00fasqueda quisimos ver si pod\u00eda traer las 4 mayores compras aceptadas que contengan la palabra "Software " en el campo del RubroN3.

Restricciones

A pesar de que se trata de un obligatorio interesante con nuevas herramientas que probar y nuevas metodologías, existen restricciones que nos enfrentamos al realizar este trabajo.

- Tiempo: Fue la principal restricción, ya que los integrantes del grupo contaban con horarios diferentes para poder trabajar en conjunto ya sea por trabajo, o estudio de otras materias, incluyendo obligatorios y parciales.
- Datos con errores: Se trabajó con datos del estado chileno para cumplir con la pauta de llegar al millón de datos requeridos para el trabajo. Debido a esto nos ocurrió que se encontraban con formato erróneo para poder trabajar sobre los mismos. Esto dificulta enormemente el proceso de Ingeniería de atributos, tomándonos la mayor cantidad de tiempo poder sobrepasar esta etapa.
- Experiencia: El equipo no cuenta con experiencia relevante en este tipo de análisis, siendo su primera aproximación al concepto práctico de Big Data y además de un informe de este estilo.

Data set depurado

<https://drive.google.com/drive/folders/1qS4Ed33NxyOUSiwmysYcoAYT1wE2zkGy?usp=sharing>

Archivos entregados

Junto a este informe se entregaron los siguientes archivos:

- Obligatorio 2022 - Ivan Monjardin y Francisco Rossi.pdf - Informe del trabajo
- Obligatorio_Francisco_Rossi_Ivan_Monjardin.ipynb - Notebook principal
- Presentacion.pdf - Presentación Power Point utilizada en la defensa
- importar-mes.ipynb - Notebook utilizado con Papermill para la importación de cada mes individual
- Graphs - Carpeta con todas las gráficas en formato .png del informe y de la presentación
- ComprasChile2021-2022.html - Reporte realizado utilizando Pandas
- ComandosSolrDocker.txt - Comandos utilizados para crear la instancia de Solr utilizando Docker

Bibliografía

Bonillo, P. (2018). Propuesta de una Arquitectura de Gestión de Grandes Volúmenes de Datos para la Analítica en Tiempo Real bajo Software Libre. Linked In, 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/>

Docuemntación de Papermill, Link: <https://papermill.readthedocs.io/en/latest/>

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

Iconos de la arquitectura, Link: <https://icon-icons.com/es/>

Definición OSCV, Link: <https://standard.open-contracting.org/latest/en/>