Toma de decisiones de inversión bursátil basadas en el seguimiento de las operaciones de compra de acciones realizadas por los directivos de empresas cotizadas en la bolsa de valores norteamericana.

(InsiderDog)

Memoria

7 de enero de 2021

Grado de Ingeniería Informática

Itinerario de desarrollo de software

Trabajo de fin de Grado

Área de Business Intelligence

Autor: Cristian Palau Tarragó (cpalaut@uoc.edu)

Director del TFG: Humberto Andrés Sanz

Responsable del Área: Atanasi Daradoumis Haralabus

FICHA DEL TRABAJO FIN DE GRADO

Título del trabajo	Toma de decisiones de inversión bursátil basadas en el seguimiento de las operaciones de compra de acciones realizadas por los directivos de empresas cotizadas en la bolsa de valores norteamericana (**InsiderDog**).
Nombre del autor	Cristian Palau Tarragó
Nombre del consultor	Humberto Andrés Sanz
Fecha de entrega	enero-2021
Área de trabajo final	Business Intelligence
Titulación	Grado en Ingeniería Informática

Resumen del trabajo

En el mundo de la inversión bursátil, no existe ningún indicador que proporcione a un inversor el momento adecuado para establecer una operación de compra con el objetivo de obtener beneficio económico y que sea 100% fiable.

Aunque esta afirmación es cierta, sí que existen indicadores que pueden proporcionar indicios de que la tendencia de un mercado bursátil o el precio por acción de una compañía cotizada va a cambiar, proporcionando una revalorización positiva en su valor.

Uno de estos indicadores se basa en el seguimiento de las operaciones de compra realizadas por los directivos de las empresas cotizadas en la bolsa de valores norteamericana, llamados *"insiders",* y que al participar en las operaciones diarias dentro de la empresa pueden conocer mejor el futuro de la misma y la evolución general del mercado.

Este proyecto propone la creación de un "*data warehouse"* que permita almacenar todas las operaciones de compra de acciones comprendidas entre los años 2003 hasta la actualidad (2020) así como el desarrollo de un software "ETL" que alimente el sistema cuando se produzcan nuevas operaciones.

El objetivo final de este proyecto es poder disponer de un sistema analítico que ofrezca la posibilidad de la realización de análisis del mercado y de las compañías cotizadas, así como la obtención patrones de comportamiento generados por estos directivos, proporcionando un indicador con un alto grado de fiabilidad para la toma de una decisión de inversión.

Palabras clave

Bolsa de valores, NYSE, NASDAQ, directivo, acciones, información privilegiada

Abstract

In the world of stock investment, there is no indicator that provides an investor with the right time to establish a purchase operation with the aim of obtaining economic profit and that is 100% reliable.

Although this statement is true, there are indicators that can provide indications that the trend of a stock market or the price per share of a listed company will change, providing a positive revaluation in its value.

One of these indicators is based on the monitoring of purchase operations carried out by managers of companies listed on the North American stock market, called *insiders*, and that by participating in daily operations within the company they can better understand the future of the company and the general evolution of the stock market.

This project proposes the development of a *data warehouse* that allows to store all the share purchase operations between the years 2003 to the present (2020) as well the development of ETL software that feeds the system when new purchase operations occur.

The final objective of this project is to be able to have an analytical system that offers the possibility of conducting market and listed company analysis, as well as obtaining behavior patterns generated by these managers, providing an indicator with a high degree of reliability for making an investment decision.

Keywords

Stock Market, NYSE, NASDAQ, managers, shares, insider trading

Índice

Ilustraciones y gráficos

1 INTRODUCCIÓN

La temática de este Trabajo de final de grado (TFG) será el análisis de las operaciones de compra realizadas por los directivos de las empresas cotizadas en la bolsa norteamericana, denominados *"insiders"* y demostrar, que a través de su seguimiento, es posible predecir futuros movimientos alcistas en el precio, tanto en un índice bursátil como en el precio de las acciones de las empresas en las que estos directivos trabajan o tienen algún tipo de participación en su capital o gestión.

Con este trabajo, se quiere demostrar que la famosa frase: "Los "*insiders"* pueden vender acciones por muchos motivos, pero si las compran solo existe uno: piensan que el precio subirá.", pronunciada por uno de los mejores inversores de la historia, Peter Lynch[1], es cierta.

Al realizar una inversión en el mercado bursátil no existe ninguna garantía de éxito pero la predicción de un movimiento al alza en el precio de una acción, puede proporcionar al inversor la ventaja de disponer de una información, que aun siendo pública, puede ser difícil de interpretar individualmente pero que analizada en su conjunto tiene un gran valor, ofreciendo un indicador fiable, que usado en conjunto con otros indicadores bursátiles, permita aumentar la probabilidad de una decisión correcta de inversión.

Para llegar a esta demostración, será necesario analizar y realizar operaciones de extracción, transformación y carga "ETL" de documentos presentados en la comisión de bolsa y valores de Estados Unidos "SEC" desde el año 2003 hasta la actualidad que contienen información detallada sobre las operaciones realizadas por los *"insiders"* y a través de un criterio estricto, saber diferenciar aquellas operaciones que aportarán valor al análisis y excluir todas aquellas cuyo valor es confuso, erróneo o simplemente no determinante. Esta información es pública y accesible por cualquiera, pero como se podrá ver, no exenta de errores que se tendrán que tener en cuenta para poder obtener la información más precisa posible.

El TFG no se limitará al análisis de las operaciones realizadas en la bolsa de valores más conocida del mundo, la de Nueva York, sino que analizará todas las operaciones de compra presentadas en los diferentes mercados que conforman todo el mundo bursátil norteamericano, como por ejemplo, el Nasdaq Stock Exchange o el OTC Market y demostrar que el comportamiento de los *"insiders"* no varía dependiendo de donde coticen las acciones de su empresa.

A través de la construcción de un *"data warehouse"*, se pretende demostrar la eficiencia de las operaciones realizadas por un grupo de inversores, los *"insiders"*, del que no se habla nunca en los medios de comunicación pero que son el verdadero "*Smart Money*" (dinero inteligente), ya que no hay mejor indicador sobre la salud y el futuro de la economía que observar las operaciones de los que, con su gestión en el día a día, son partícipes directos de ella.

[1] Adam Barabone, «What Investors Can Learn From Insider Trading»,2020, https://www.investopedia.com/articles/02/061202.asp

1.1 CONTEXTO Y JUSTIFICACIÓN

Según estudios[2] realizados, el porcentaje de éxito en las inversiones bursátiles realizadas por inversores norteamericanos es menor del 10%. Con unas cifras de éxito tan pobres, es difícil creer que exista un grupo de inversores que puedan superar estas cifras de forma regular y que pasen tan desapercibidos en una cultura de inversión bursátil como es la norteamericana. Este grupo de inversores, los *"insiders",* son los directivos que trabajan o tienen alguna participación en el capital o en la gestión de una empresa cotizada en el mercado bursátil.

El "*Insider Trading"* es la negociación de productos financieros de una empresa, como por ejemplo, acciones, realizada en posesión de material público o no público de la misma y cuya negociación es realizada por directivos o entidades con intereses en la empresa.

Las tres piezas de información básica en una operación de "*Insider Trading"* son: el "*issuer"*, que es la compañía cotizada sobre la que se realiza la transacción, el "*report owner"*, que es el "*insider"* que la realiza y el detalle de las transacciones realizadas, ya sean operaciones de compra, venta u otro tipo, como por ejemplo, la ejecución de una opción de compra de acciones.

Nejat Seyut, profesor e investigador de la universidad de Michigan en Estados Unidos y experto en el campo del "*Insider Trading",* a través de las investigaciones recogidas en su libro[3], descubrió que cuando los *"insiders"* compran acciones de sus propias empresas, la rentabilidad de esas acciones supera la rentabilidad del mercado en un 8,9% durante los siguientes doce meses.

Lo primero que llama la atención del "*Insider Trading"* es la definición "material no público" que puede llevar a pensar que esta negociación puede ser ilegal. Lo es y no lo es. Para afirmar que alguien ha cometido un acto ilícito hay que demostrarlo y para ello el organismo regulador encargado de controlar estas operaciones es la "*Securities Exchange Comisión (*SEC)" cuyo principal objetivo es velar por los intereses de los inversores. Demostrar que una operación de "*Insider Trading"* es ilegal puede llegar a ser muy complejo ya que algunas compañías están formadas por cientos de *"insiders"* y es complicado poder monitorizar todas las transacciones y detectar posibles casos de fraude.

Un ejemplo de "*Insider Trading"* ilegal ocurre cuando el hijo de un "*insider",* que trabaja en una empresa cotizada, escucha comentar a su padre en una conversación telefónica que su empresa va a conseguir un contrato gubernamental de cientos de millones de dólares y compra para sí mismo o a través de un tercero de confianza, acciones de esa empresa antes de que la información se haga pública al resto de inversores con el objetivo de obtener una rentabilidad ilícita.

[2] Roberts, L. «Americans are still terrible at investing, annual study once again shows» https://www.marketwatch.com/story/americans-are-still-terrible-at-investing-annual-study-once-again-shows-2017-10-19

[3] *Seyut, Nejat,* «*Investment Intelligence from Insider Trading*. (2000)»: https://www.amazon.com/Investment-Intelligence-Insider-Trading-Press/dp/0262692341

Un caso real de "*Insider Trading"* ilegal, que generó mucho ruido mediático fue la venta ilegal de acciones de la empresa ImClone Systems[4] por parte presentadora televisiva norteamericana Martha Stewart que fue acusada de utilizar información privilegiada, proporcionada por su íntimo amigo y CEO de la empresa, Samuel D. Waksal. La presentadora, que fue condenada por este hecho, se aprovechó de la información que le proporcionó Waksal para vender acciones de la compañía un día antes de que esta anunciara que uno de sus futuros medicamentos estrella, "*Erbitux",* no había pasado la aprobación del organismo regulador de medicamentos norteamericano (FDA).

Por casos como este y muchos otros que se han producido en el pasado, la "SEC" ha ido modificando las leyes respecto al "*Insider Trading"* durante las últimas décadas. La última modificación se produjo en el año 2002, a raíz de los escándalos bursátiles de los años 2000, como el fraude de Enron Corporation, creando una nueva ley denominada "***Sarbanes-oxley act"*** [5], en la que definieron una medida concreta para el "*Insider Trading"* que reduce el número máximo de días en los que los *"insiders"* pueden notificar una transacción realizada, pasando de diez a dos días hábiles después de haberse producido.

Estas regulaciones, al contrario de lo que se podría pensar, no han reducido el número de operaciones realizadas por este grupo de inversores. De hecho, el avance de las nuevas tecnologías y en concreto la publicación electrónica de los documentos en la "SEC", ha permitido no sólo facilitar y acelerar la publicación de esta información, sino también mejorar la transparencia de estas operaciones ante los accionistas de la empresa.

El "*Insider Trading"* se practica en diferentes países, pero es en Estados Unidos donde se producen las mejores condiciones para realizar un análisis detallado, ya que la "SEC" publica, de forma diaria a través de sus sistemas informáticos, unos documentos llamados *"form 4 filings"* que los *"insiders"* van enviando al organismo regulador para informar de sus operaciones. Estos *"filings"* contienen toda la información detallada de las transacciones, así como datos sobre el rol del insider dentro de la compañía, muy útiles para poder excluir del análisis todos aquellos roles, como por ejemplo, las grandes participaciones accionariales.

A pesar de que los datos sobre estas operaciones son públicos, es difícil encontrar información o cobertura sobre el "*Insider Trading"* en los medios escritos o digitales especializados del mundo financiero. La información que podemos encontrar en estos medios es la misma que está disponible en el sitio Web de la "SEC", limitándose a informar de la publicación de un *"form 4 filing"* sin ningún tipo de análisis o interpretación.

Al no existir análisis detallados sobre el seguimiento de las operaciones realizadas por los *"insiders"* de forma diaria, es necesaria la creación de un *"data warehouse"* que permita almacenar toda esta información, para que posteriormente sea explorada, tratando de encontrar evidencias de que a través del seguimiento de las

[4] Wikipedia: «*Imclone stock trading case*.» (2020). https://en.wikipedia.org/wiki/ImClone_stock_trading_case

[5] Bill, K. «*Sarbanes Oxley Act*. »: https://www.investopedia.com/terms/s/sarbanesoxleyact.asp

operaciones de compra, es posible predecir un futuro cambio en el precio de una acción o índice bursátil.

1.2 OBJETIVOS

El objetivo de este TFG es realizar un análisis exploratorio de los datos generados por las operaciones de compra de acciones realizadas por los *"insiders"* y obtener unos resultados que puedan usarse como indicador para la predicción del futuro de los precios.

Para poder llegar a este objetivo se descargarán, procesaran y transformaran más de 3 millones de documentos en formato "XML" generados entre los años 2003 y 2020, proporcionados por la "SEC" a través de su sitio Web, utilizando una herramienta que será desarrollada para la descarga de este histórico de documentos, así como los que se vayan publicando en tiempo real a través de la tecnología "RSS" o en el momento del cierre diario del mercado. Estos documentos se almacenarán en una base de datos relacional que será utilizada como *"data warehouse"*. Antes de proceder a su almacenamiento será necesario hacer operaciones de "*data wrangling"* en algunos datos para su posterior importación al modelo relacional con el fin de poder realizar un análisis más preciso.

Una vez almacenados en el *"data warehouse"* y usando software con el soporte de una librería para el análisis de datos, se ofrecerá una conclusión final.

Esta conclusión final se dividirá en 2 apartados: **análisis del mercado global** y **análisis de compañía cotizada**.

El análisis del mercado global ***"Daily Filing Purchases"*** tratará de demostrar como los "*insiders*" pueden predecir, con cierta antelación, fuertes movimientos alcistas a través de 3 ejemplos: el final de la crisis financiera (2007-2009)[6], el "*crash*" del "*Black Monday*" (2011)[7] y el "*crash*" de la pandemia del "covid-19" (2020)[8].

El análisis de compañía cotizada, tratará de demostrar como los *"insiders"* forman parte del denominado "*Smart Money*" (dinero inteligente) y se analizarán las operaciones de compra en una misma compañía "***Buying Cluster***", realizadas por dos o más *"insiders"* y en un periodo máximo de treinta días entre todas ellas, las operaciones de compra realizadas por un "*insider"* que ha acertado en el pasado realizando compras en diferentes compañías "***IQ buy***" y las operaciones de compra realizadas por *"insiders"* cuyo rol ejecutivo dentro de la compañía es clave en la gestión de la operativa en el día a día, como por ejemplo, el CEO ("Chief executive officer") o el CFO ("*Chief Financial Officer*"), denominadas "***CEO/CFO 100K buy***".

[6] Wikipedia. (2020). «*Bear Market (2008-2009)*». https://en.wikipedia.org/wiki/United_States_bear_market_of_2007

[7] Wikipedia. (2020). «*Black Monday (2011)*». https://en.wikipedia.org/wiki/Black_Monday_(2011)

[8] Wikipedia. (2020). «*Stock Market Crash (2020)*». https://en.wikipedia.org/wiki/2020_stock_market_crash

Queda fuera del alcance de este trabajo, analizar otras operaciones, como las de venta, ya que como se verá durante el análisis, no son concluyentes para realizar una predicción sobre la evolución de los precios y que para que pudieran ofrecer información de valor, sería necesario aplicar modelos predictivos basados en "*machine learning*".

1.3 ENFOQUE Y METODOLOGÍA SEGUIDA

Este TFG se basará en el uso de dos metodologías muy utilizadas en el ámbito de la gestión de proyectos y el desarrollo de software.

Para la gestión del proyecto, se usará la metodología propuesta por el "*Project Managament Institute*" a través de su guía **PMBOK**[9] ("*Project Management Body of Knowledge*").

Para el proceso lógico del Desarrollo de Software, se usará la metodología SDLC ("*System Development Life Cycle*"), en concreto su variante de desarrollo secuencial, "***Waterfall***", una de las más utilizadas en el área del desarrollo de proyectos de "*data warehouse*" en los últimos años. (Rainardi, V. (2008). *Building a Data Warehouse*. Capítulo 3, página 49).

Para la gestión del proyecto, se usarán los macroprocesos expuestos en la ilustración 1, con alguna pequeña variación.

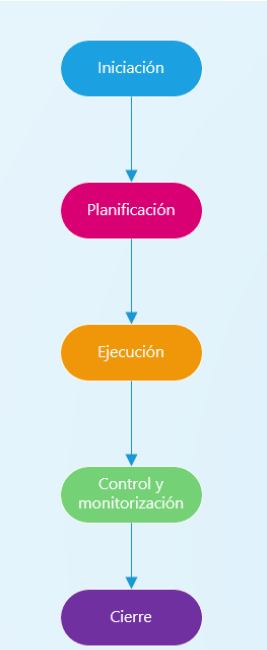


Ilustración 1 Macroprocesos de la guía PMBOK

La metodología "SDLC Waterfall" se basa en los siguientes pasos:

- **Estudio de viabilidad**
- **Requerimientos**
- **Arquitectura**
- **Diseño**
- **Desarrollo**
- **Pruebas**
- **Despliegue**
- **Soporte**

Como se puede observar, el estudio de viabilidad, el despliegue, y el soporte del sistema son tres pasos que deben adaptarse por tratarse de un proyecto académico.

[9] García, L. A. (2016). «TFC. Gestión de proyectos». http://openaccess.uoc.edu/webapps/o2/bitstream/10609/45590/7/lameijideTFC0116memoria.pdf

El producto final requiere un mantenimiento que deberá proporcionarse una vez concluido el proyecto. También quedará fuera del ámbito del TFG, aquella documentación adicional, como por ejemplo, el manual del usuario o la guía de despliegue en un entorno virtual.

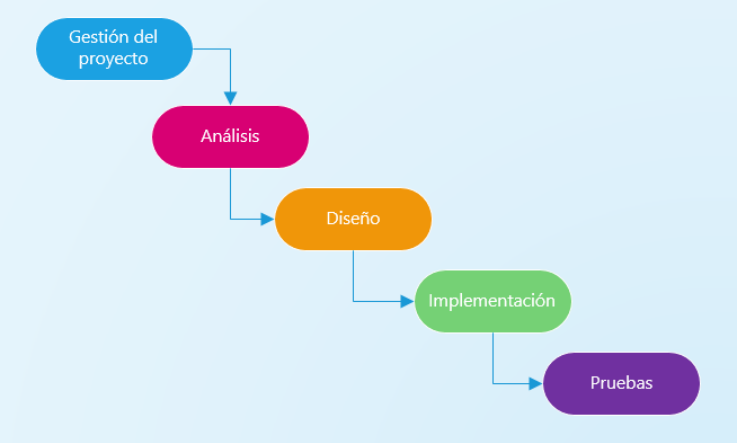


Ilustración 2 Estructura de desglose de trabajo "EDT"

Al tratarse de un producto de "*reporting*", a través de un *"data warehouse"* y al haber escogido una metodología SDLC secuencial, será necesario establecer un seguimiento a través de tres informes de proyecto en el que las tareas serán revisadas y corregidas durante todo el ciclo de vida del desarrollo.

Para lograr el objetivo del proyecto y crear los productos entregables requeridos, se ha optado por definir un desglose de trabajo "EDT", tal como sugiere el **PMBOK**, mostrado en la ilustración 2, con el objetivo de adaptar las dos metodologías.

1.4 PLANIFICACIÓN DEL TRABAJO

La planificación del trabajo establece las tareas principales y los entregables así como aquellas tareas adicionales necesarias para el desarrollo del proyecto teniendo en cuenta los riesgos a los que se enfrenta el proyecto así como las correcciones necesarias para su correcta finalización.

El proyecto requerirá un total de 330 horas de trabajo. Se deben tener en cuenta los siguientes puntos que condicionan el desarrollo del proyecto:

- Existe sólo un recurso humano para el desarrollo del proyecto que deberá ejercer los roles de Jefe de Proyecto, Analista Programador, Administrador de Base de Datos y Analista de Datos.

- La dedicación diaria de este recurso será de 3-4 horas, pudiendo ampliarse o reducirse este número de horas dependiendo de la evolución del proyecto y las posibles desviaciones producidas.

- Todo el desarrollo se realizará en un ordenador personal PC con las siguientes características:

  - **Procesador i7-5820 con 6 núcleos y 12 hilos.**
  - **64 GB de RAM**
  - **Disco duro SSD de 500 GB**

- El proyecto utilizará entre otras, las siguientes tecnologías principales:
  - **Windows 10 con uso del sistema WSL-2**
  - **Python 3.8 como lenguaje de programación**
  - **MariaDB 10.5 como base de datos relacional**
  - **PyCharm 2020.1.1 como IDE de programación en Python**
  - **Pandas y SQLAlchemy como librerías principales para el análisis de datos y la persistencia de datos.**

- El proyecto se desarrollará en su totalidad entre las fechas del 21/09/2020 y el 07/01/2021. El total de horas de desarrollo del proyecto es de 330. De las cuales 310 son para desarrollo del proyecto y 20 para posibles desviaciones.

- La finalización del proyecto sin la entrega de la memoria final está prevista para el 09/12/2020, momento en el que se entregará el segundo informe de seguimiento (PAC3).

1.4.1 VALORACIÓN ECONÓMICA

Como se ha comentado en el apartado anterior, el recurso que desarrollará el proyecto ejercerá cuatro roles durante las diferentes fases de su desarrollo.

A continuación se detallan las horas y los costes de este recurso en sus diferentes roles:

ROL	Horas	Coste/Hora €	Total €
Jefe de Proyecto	85	70	5.950
Administrador de Base de Datos	34	35	1.190
Analista programador	166	45	7.470
Analista de datos	25	52	1.300
Total			**15.910**

También hay que tener en cuenta, que existe el siguiente coste adicional:

Concepto	Servidores	Horas	Total horas	Coste/Hora €	Total €

VULTR SSD Cloud Instances	17	216	3.672	0,005	18,36

El total del coste del proyecto asciende a: **15.928,36 €**

1.4.2 ANÁLISIS DE RIESGOS

A continuación, se detallan los riesgos que se contemplan en el desarrollo del proyecto y que pueden condicionar su finalización así como las acciones correctivas que se deberían realizar para la gestión de los mismos.

ID	Riesgo	Descripción	Probabilidad	Impacto
R01	Diseño incorrecto del modelo de datos	Si el modelo de datos no es correcto puede conllevar errores en la interpretación de los resultados finales.	Media	Alto
R02	Problemas de hardware	Daños físicos en el equipo de desarrollo.	Media	Alto
R03	Desviaciones	Errores en la estimación del coste de las tareas.	Baja	Alto
R04	Problemas de software	Problemas de instalación, configuración y uso del software.	Media	Alto
R05	Problemas laborales	Tener que dedicar más horas a temas laborales o la posibilidad de viajar.	Baja	Alto
R06	Disponibilidad de la información de terceros	Problemas en la descarga de la información pública de la SEC.gov	Baja	Alto
R07	Falta de documentación	No disponer de la documentación necesaria para poder desarrollar el proyecto.	Baja	Medio

Para cada uno de los riesgos detectados se plantean las siguientes acciones correctivas:

ID	Riesgo	Acción correctiva	Coste
A01	**R01**	Rediseñar el modelo.	Alto
A02	**R02**	Sustitución del hardware.	Alto
A03	**R03**	Replanificación de las tareas.	Medio
A04	**R04**	Buscar documentación o ayuda de un tercero .	Medio
A05	**R05**	Invertir más tiempo en los días festivos y en horas muertas durante los viajes.	Bajo

A06	**R06**	Utilizar mecanismos alternativos como el uso de DVD de datos vendidos por terceros.	Medio
A07	**R07**	Búsqueda de información en los foros de Internet.	Medio

1.4.3 ACTIVIDADES Y TAREAS DEL PROYECTO

Las actividades y el conjunto de tareas de cada una de ellas se detallan a continuación, incluyendo la fecha de finalización y el total estimado de horas para su finalización.

	Nombre	Fecha de finalización	Coste en horas
0	**Gestión del TFG**	09/12/2020	**100**
1	**Propuesta**	**12/10/2020**	**44**
2	Análisis inicial	25/09/2020	10
3	Estimación de costes	27/09/2020	4
4	Definición de objetivos	02/10/2020	10
5	Documentación y entrega PAC1	12/10/2020	20
6	**Planificación**	**12/10/2020**	**44**
7	Establecer tareas	25/09/2020	10
8	Planificación de tareas	28/09/2020	6
9	Definición de hitos	03/10/2020	10
10	Análisis de riesgos y correcciones	06/10/2020	6
11	Documentación inicial	12/10/2020	12
12	**Seguimiento**	**09/12/2020**	**12**
13	Elaboración del plan de trabajo (PAC1)	12/10/2020	4
14	Elaboración de PAC2	09/11/2020	4
15	Elaboración de PAC3	09/12/2020	4
16	**Ejecución del TFG**	09/12/2020	**210**
17	**Análisis**	**23/10/2020**	**26**
18	Requisitos funcionales y no funcionales	17/10/2020	10
19	Análisis funcional y de modelo de datos	21/10/2020	10
20	Análisis de diseño de arquitectura	23/10/2020	6
21	**Diseño**	**09/11/2020**	**33**
22	Diseño conceptual	31/10/2020	6
23	Diseño lógico de la base de datos	02/11/2020	8
24	Diseño físico de la base de datos	04/11/2020	9
25	Documentación PAC2	09/11/2020	10
26	**Implementación**	**24/11/2020**	**90**
27	Instalación y configuración del entorno	10/11/2020	2
28	Descarga de *"form 4" "filings"*	19/11/2020	20

29	Desarrollo del software "ETL"	20/11/2020	48
30	Importación de datos	24/11/2020	12
31	Documentación PAC3	24/11/2020	8
32	**Pruebas**	09/12/2020	**61**
33	Análisis de datos	30/11/2020	27
34	Conclusiones del análisis de datos	08/12/2020	20
35	Documentación PAC3	09/12/2020	14

1.4.4 HITOS DEL PROYECTO

Para el proyecto se han definido los siguientes hitos en los que se incluyen los documentos de seguimiento que hay que entregar periódicamente (PAC) así como todos aquellos que son necesarios para el correcto seguimiento del proyecto, ayudando a controlar su evolución y a corregir posibles desviaciones.

Nombre	Fecha del hito	Tipo
Propuesta inicial aceptada	**21/09/2020**	**EJECUCIÓN**
Planificación cerrada	**12/10/2020**	**EJECUCIÓN**
Análisis cerrado	**21/09/2020**	**EJECUCIÓN**
Diseño cerrado y entrega PAC2	**05/11/2020**	**SEGUIMIENTO**
Descarga *"form 4 filings"* completa	**19/11/2020**	**EJECUCIÓN**
Desarrollo "ETL" completo	**20/11/2020**	**EJECUCIÓN**
Importación de datos completada	**24/11/2020**	**EJECUCIÓN**
Conclusiones del análisis de datos	**08/12/2020**	**EJECUCIÓN**
Entrega de PAC3	**09/12/2020**	**SEGUIMIENTO**
Entrega de memoria	**07/01/2021**	**EJECUCIÓN**

Aunque dentro del proyecto, todos los hitos son importantes, hay que destacar que los siguientes hitos son esenciales para la correcta finalización del proyecto:

- **Descarga *"form 4 filings"* completa**
- **Desarrollo "ETL" completo**
- **Importación de datos completada**
- **Conclusiones del análisis de datos**

Dada su importancia, se llevará un estricto seguimiento sobre el cumplimiento de las fechas de entrega.

1.4.5 DIAGRAMA DE GANTT

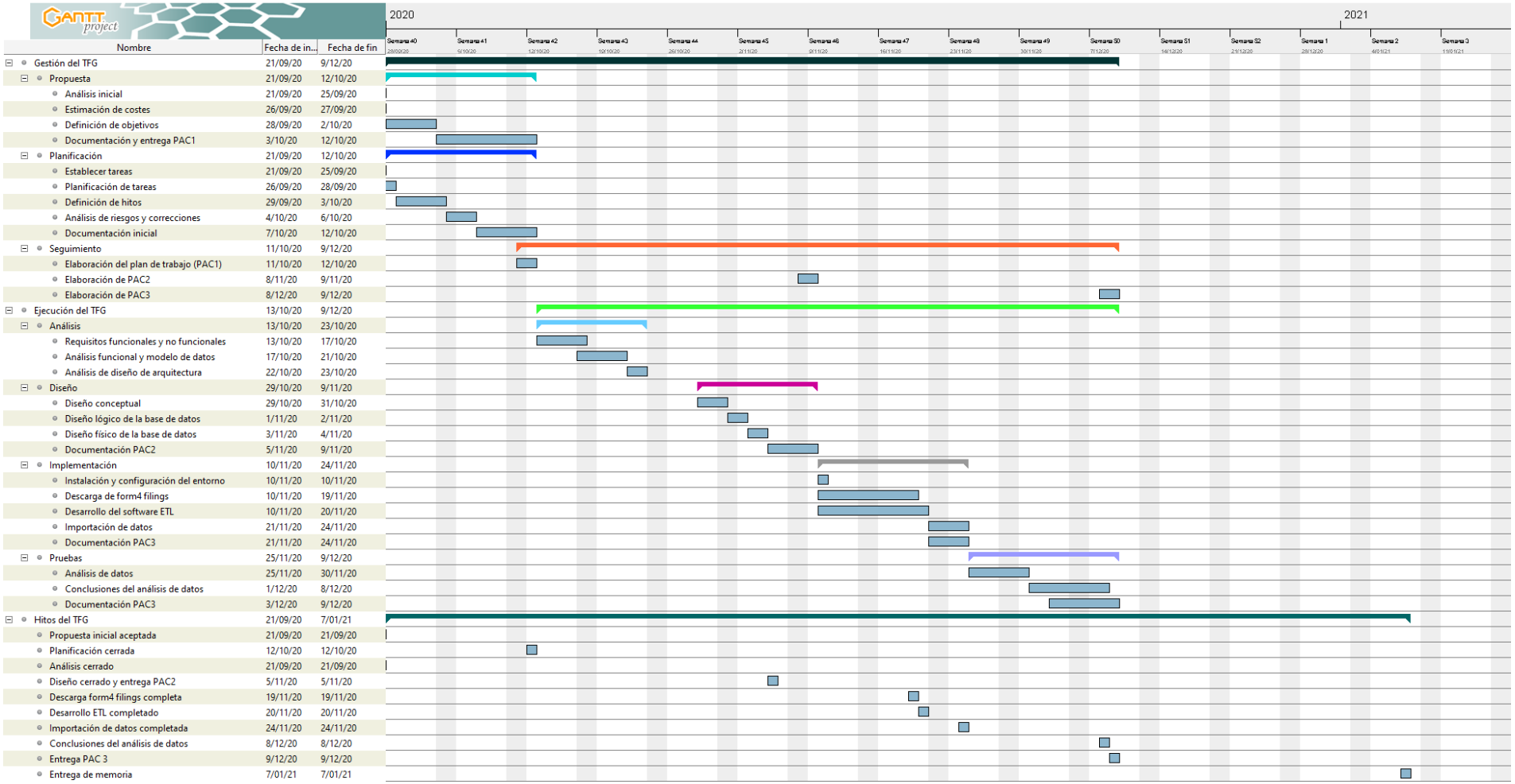


Ilustración 3 Diagrama de Gantt

1.4.6 PRODUCTOS QUE SE VAN A OBTENER

Durante las diferentes fases del proyecto se van a obtener los siguientes entregables:

- **Análisis de los requisitos funcionales y no funcionales**
  - Descripción de todos los servicios y actividades que el sistema proporciona así como sus características, limitaciones y prestaciones.

- **Análisis del diseño conceptual del "*data warehouse*" y del software "ETL"**
  - Descripción del modelo de dominio para el desarrollo del *"data warehouse"* necesario para el almacenamiento así como el diseño conceptual del software "ETL" que lo va a alimentar.

- **Análisis del diseño físico de la base de datos relacional**
  - Descripción del modelo de datos y del diseño físico de la base de datos relacional que fundamenta el *"data warehouse"*.

- **Conclusiones del análisis del análisis exploratorio, de mercado y compañía.**
  - Documento de conclusiones sobre el análisis exploratorio, de mercado y de compañía extraído del *"data warehouse",* objetivo final del proyecto.

- **Aplicación "ETL"**
  - Software desarrollado en el lenguaje de programación Python para las operaciones "ETL" del proyecto.

- **"form 4 filings"**
  - Archivo en formato ZIP que incluye todos los "*form 4 filings*" desde 2003 a la fecha de cierre del proyecto.

- **Base de datos**
  - *Backup* de la base de datos relacional MariaDB comprimida en formato ZIP.

- **Documento de configuración e instalación del proyecto**
  - Documentación de como configurar e instalar el proyecto en un entorno local.

2 ANÁLISIS DEL SISTEMA

El proyecto se basa en desarrollar un sistema *"data warehouse"* que permita almacenar en una base de datos relacional todos los documentos, denominados *"form 4 filings",* presentados de forma telemática en la "*Securites Exchange Commission* SEC", sobre operaciones de "*Insider trading"*.

Aunque en el proyecto se almacenarán todos los documentos presentados desde el año 2003 hasta la actualidad, el objetivo principal, será el análisis de las operaciones de compra de acciones por parte de los *"insiders"*.

Este *"data warehouse"*, será constantemente alimentando con información procedente de la "SEC" y será el punto de partida para la generación de un análisis posterior con el fin de encontrar patrones de comportamiento que puedan predecir movimientos futuros al alza en el precio de un índice bursátil o una empresa cotizada.

Uno de los pilares básicos del proyecto es el desarrollo de una herramienta "ETL" que permita, la descarga, el tratamiento y la persistencia de estos documentos en el *"data warehouse"*.

Por todo ello, debemos primeramente analizar el dominio del problema para poder presentar el análisis de requisitos, el catálogo de requisitos funcionales y no funcionales así como el análisis del modelo de datos y el diseño conceptual que vamos a utilizar.

2.1 ANÁLISIS DE DOMINIO

En el año 1983, la "SEC" empezó a desarrollar el sistema "**EDGAR**"[10] con el principal objetivo de obligar a las empresas que cotizan en los diferentes mercados, la presentación telemática de la mayoría de documentos públicos que exigen las leyes del mercado bursátil.

No fue, hasta el año 1996, cuando una vez completado el desarrollo de "EDGAR", se impuso la obligación real de la presentación telemática, exigiendo a todas las empresas la modificación de sus sistemas informáticos con el objetivo de la presentación telemática de los documentos, siendo este el canal único y obligatorio para dichas presentaciones, permitiendo al regulador una mayor eficiencia en el control y seguimiento de las operaciones bursátiles.

El sistema "EDGAR", que sigue vigente en la actualidad, recibe más de tres mil documentos al día, de los cuales, los denominados *"form 4"* o "*Statement of Changes in*

[10] Wikipedia, «"EDGAR" System»,2020, https://en.wikipedia.org/wiki/"EDGAR"

Beneficial Ownership Overview", son los que contienen la información sobre operaciones relacionadas con "*Insider trading"*.

Desde el año 1996, ha sido posible para un inversor privado descargar los documentos *"form 4"* pero la información disponible era limitada, incompleta y en un formato no estructurado, imposible de interpretar usando un ordenador.

Fue a mediados del año 2003 cuando la "SEC" empezó a desarrollar un sistema para que el público general pudiera descargar la información en formato **"XML"** que es un formato semi estructurado y por lo tanto abriendo la posibilidad de usar ordenadores para su proceso.

La definición de un esquema "**XSD"**, permitió establecer una descripción de los elementos que debía contener el "XML", especificando una serie de reglas que para que un documento "XML" se considerara válido y fuese aceptado por el nuevo sistema.

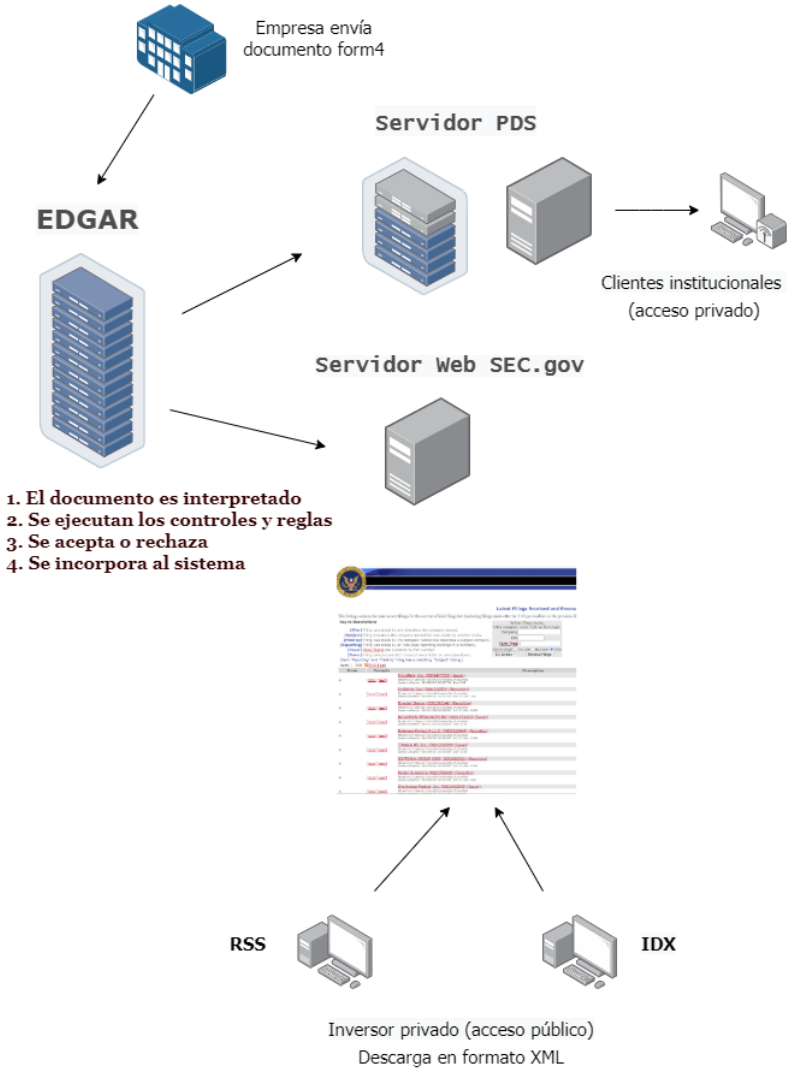


Ilustración 4 Funcionamiento del sistema "EDGAR"

Como se puede ver en la ilustración 4, las empresas utilizan “EDGAR” para enviar los documentos y en el caso de los *“form 4”*, los documentos pasan unas validaciones extra proporcionadas por el propio formato “XML”, como por ejemplo, que existan los datos del “*issuer”* que genera el documento y del “*insider”* que realiza la transacción, así como la transacción en sí, que son necesarios para dar validez al documento. Podemos observar que cada documento remitido por estas empresas pasa una serie de controles y que puede ser rechazado. Todo este proceso se hace en tiempo real y esta automatizado para que la información esté disponible lo más rápido posible tanto para los inversores institucionales como para los inversores privados.

También podemos ver que existe un servicio institucional de pago, llamado “PDS”, que ofrece a los inversores institucionales, como por ejemplo, los gestores de fondos de inversión, herramientas adicionales que permiten analizar en detalle todos los documentos enviados que han sido procesados, clasificados y validados.

El inversor privado, al que va enfocado este proyecto, puede descargar los documentos *“form 4”* de dos formas distintas. La primera es utilizando un servicio “**RSS”** (“*Really Simple Sindication”*) y la segunda es usando el servicio “**IDX”** formado por un conjunto de ficheros proporcionados por la “SEC” a través de su sitio Web que contienen el índice de todos los documentos, sean del tipo *“form 4”* o no, publicados durante un periodo determinado, que puede ser diario, trimestral o anual. La principal diferencia entre ambos sistemas es que la descarga a través de “RSS” es en tiempo real y en los ficheros “IDX” la primera disponibilidad es al final de la sesión bursátil de un día concreto.

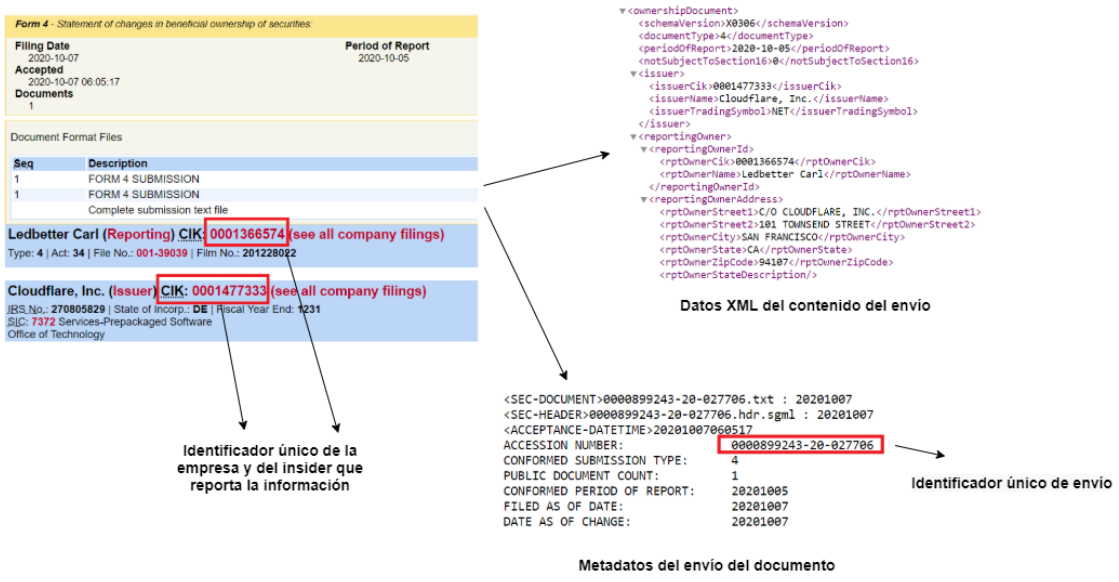


Ilustración 5 Composición de un envío “form 4”

En la ilustración 5, podemos ver el contenido de un envío que se basa en la presentación de un documento que contiene metadatos para su clasificación en el sistema “EDGAR”, tales como el identificador único dentro del sistema de la empresa que lo remite y los *“insiders”* a los que afecta (“**CIK”** o “*Central Index Key”*) así como el

documento “XML” que contiene los datos concretos de la operación que se quiere hacer pública. Una vez presentado y validado, “EDGAR” le asigna un identificador único de envío denominado “***accession number”*** que identifica al documento dentro del sistema.

Como veremos a lo largo del proyecto, la definición de este “XML”, hace posible que este proyecto pueda ser desarrollado, aunque la existencia de una validación “XSD” no evita que los datos publicados contengan errores.

Una vez analizado el funcionamiento del sistema “EDGAR”, debemos definir cómo será el proceso de “d*ata warehousing”* que se va a desarrollar para almacenar la información y que permitirá realizar el análisis final.

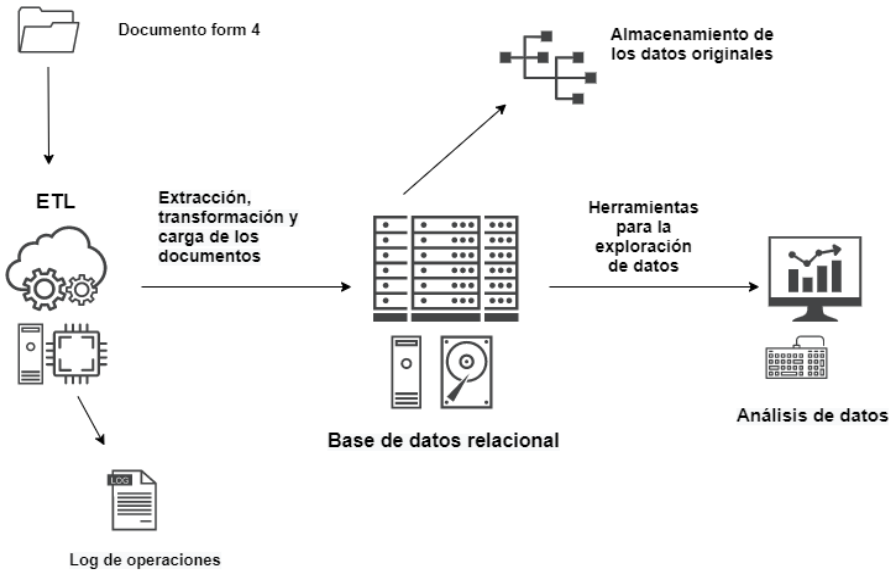


Ilustración 6 Proceso de “data warehousing”

Como se puede ver en la ilustración 6, el proceso de “*data warehousing”* se basará en un software “ETL” que será el encargado de extraer, transformar y cargar la información y en una base de datos relacional que será la responsable de almacenar toda la información.

El software tendrá la capacidad de descargar los documentos *“form 4”* del sistema “EDGAR” utilizando los servicios de descarga “RSS” o “IDX”. También tendrá la capacidad de realizar transformaciones de la información publicada y proporcionará las operaciones de persistencia necesarias para que la información quede almacenada en una base de datos relacional de forma normalizada.

Hay que destacar, que dada la naturaleza académica del proyecto y la complejidad del mismo, se deja para una futura fase, la creación de un “*Backoffice”* de apoyo al *“data warehouse”,* que ofrezca la posibilidad de establecer reglas de validación de la información, la creación de alertas personalizadas en caso de que ocurran ciertos eventos o la importación de los precios diarios de todas las acciones de las compañías cotizadas, todo ello con el objetivo de potenciar el sistema analítico.

2.2 ANÁLISIS DE REQUISITOS

2.2.1 CATÁLOGO DE REQUISITOS

El catálogo de requisitos que se ha definido tiene como objetivo exponer de forma clara, precisa, completa y verificable todas las funcionalidades y restricciones del sistema a través de los requisitos funcionales y no funcionales.

2.2.1.1 REQUISITOS FUNCIONALES

Los requisitos que detallan la funcionalidad del sistema serán los siguientes:

RF01	Descarga de los documentos "form 4"
Descripción	El sistema proporcionará un mecanismo para la descarga de documentos "*form 4*" del sitio web de la "SEC" (SEC.gov).
Importancia	Muy alta
Comentarios	Se desarrollará una herramienta que permitirá descargar los documentos *"form 4"*, ya sea en tiempo real utilizando el sistema "RSS" que proporciona la "SEC" a través de su sitio Web o de forma diaria, trimestral o anual utilizando el sistema "EDGAR IDX".

RF02	Extracción de la información en los documentos "form 4"
Descripción	El sistema permitirá extraer la información de los documentos.
Importancia	Muy alta
Comentarios	Se desarrollará una herramienta que permitirá extraer la información de los documentos descargados en formato semi estructurado "XML". Se va a extraer toda la información de los documentos según reglas específicas.

RF03	Transformación de la información en los documentos "form 4"
Descripción	El sistema permitirá transformar la información de los documentos procesados.
Importancia	Muy alta
Comentarios	Se desarrollará una herramienta que permitirá transformar la información procesada para poder adaptarla a una base de datos relacional con objetivos analíticos *"data warehouse"*.

RF04	Validación de los documentos “form 4”
Descripción	El sistema permitirá la validación de los documentos transformados.
Importancia	Alta
Comentarios	Se desarrollará una herramienta que permitirá la validación del “XML” usando un esquema “XSD” así como los controles necesarios para evitar la duplicación de datos en el sistema.

RF05	Carga de documentos “form 4” en el sistema
Descripción	El sistema permitirá cargar la información de los documentos transformados.
Importancia	Muy alta
Comentarios	Se desarrollará una herramienta que permitirá la carga de la información procesada en una base de datos relacional de forma normalizada y preparada para su uso en procesos analíticos.

RF06	Registro de operaciones
Descripción	El sistema permitirá almacenar un registro de las operaciones realizadas.
Importancia	Alta
Comentarios	Se desarrollará una herramienta que permitirá almacenar todas las operaciones que el software “ETL” realice en el sistema, tales como los documentos que se han procesados, los que no han podido validarse o los documentos incorrectos.

RF07	Análisis de datos
Descripción	El sistema permitirá analizar los datos para la extracción de conclusiones.
Importancia	Alta
Comentarios	Se desarrollará una base de datos relacional, que apoyada con otras herramientas de análisis, permitirá realizar análisis exploratorios en los datos del *“data warehouse”*, así como la creación de herramientas complementarias para detectar eventos como los “***Buying Cluster***” o “***IQ Buy***”

RF08	Alertas sobre eventos
Descripción	El sistema permitirá alertar de eventos que se produzcan.
Importancia	Alta
Comentarios	Se desarrollará una herramienta que permita enviar alertas personalizadas, ya sea por e-mail u otro canal, que alerten de eventos inusuales, como por ejemplo, los “***Buying Cluster***” o “***IQ Buy***”. *

* Este requisito será desarrollado en una futura actualización del proyecto dado su alto requisito en horas de desarrollo

2.2.1.2 REQUISITOS NO FUNCIONALES

Los requisitos detallan las restricciones y características generales del sistema serán los siguientes:

RNF01	Uso de tecnología de código abierto
Descripción	El sistema hará uso únicamente de tecnología de código abierto.
Importancia	Muy alta
Comentarios	Se utilizará el lenguaje de programación Python, la base de datos relacional MariaDB y la librería Pandas para el análisis de datos. Todos estos productos están licenciados bajo licencias de código abierto.

RNF02	Eficiencia para el análisis de datos
Descripción	El sistema ha de ser capaz de operar adecuadamente en sesiones concurrentes.
Importancia	Alta
Comentarios	El uso de la tecnología MariaDB, permite la definición de un "*data warehouse*" que puede escalar en el momento que sea preciso a través de una solución basada escalado horizontal.

RNF03	Independencia de plataforma
Descripción	El sistema ha de ser capaz de poder ser instalado y ejecutado en diferentes sistemas operativos.
Importancia	Alta
Comentarios	El sistema se desarrollará con independencia del sistema operativo sobre el que se ejecute ya sea en versiones de Microsoft Windows 10 o posteriores así como en un sistema operativo basado en UNIX.

2.2.2 ANÁLISIS FUNCIONAL

En el "*data warehouse*" podemos encontrar los siguientes actores:

- **Administrador:** Encargado de realizar las tareas a través de la ejecución del software "ETL" y del mantenimiento del sistema.
- **Analista de datos**: Encargado de extraer la información para su análisis.
- **Sistema:** El propio sistema como actor para la realización de tareas de forma autónoma.

Para capturar los requisitos funcionales se han definido los siguientes casos de uso:

CAU1	Descarga de documentos “*form 4*”

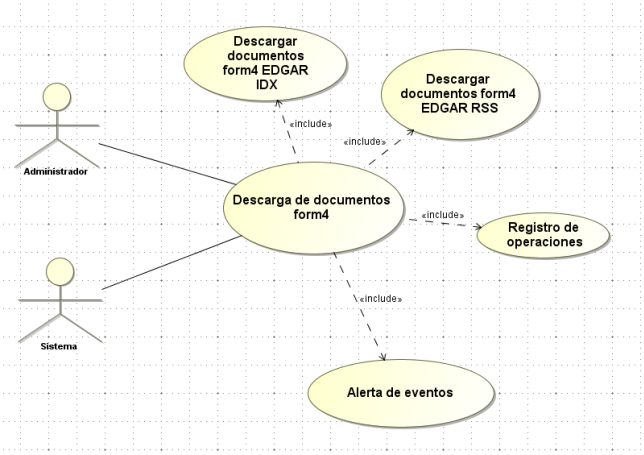


Ilustración 7 Caso de uso Descarga documentos “form 4”

CAU2	Extracción de información de documentos “*form 4*”

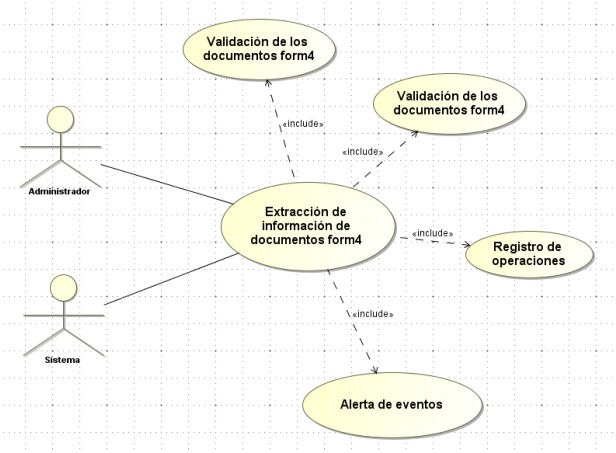


Ilustración 8 Caso de uso Extracción de información de documentos “form 4”

CAU3	Transformación de la información de los documentos “*form 4*”

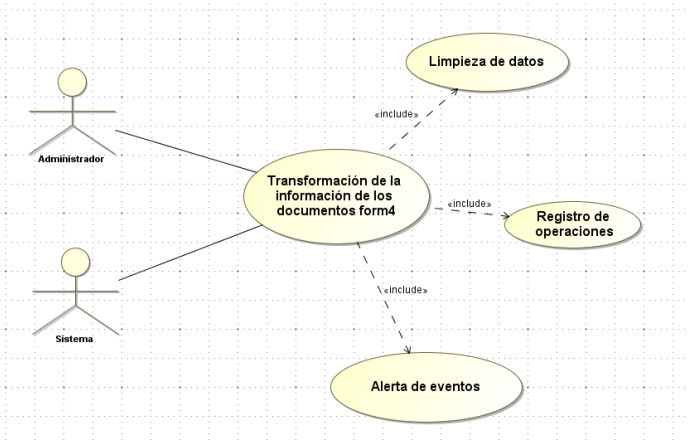


Ilustración 9 Caso de uso Transformación de la información de los documentos “form 4”

CAU4	Carga de documentos "*form 4*" en el sistema

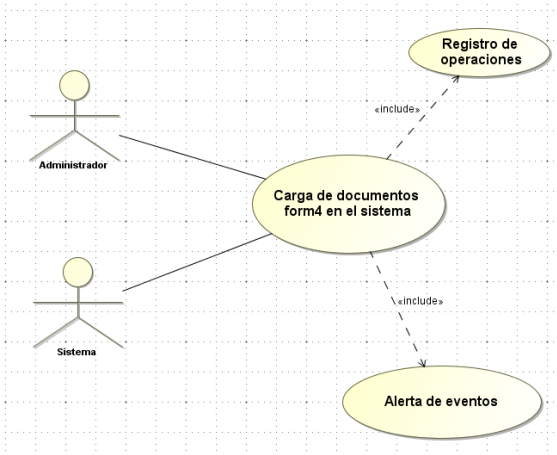


Ilustración 10 Caso de uso Carga de documentos "form 4" en el sistema

CAU5	Validación de documentos "*form 4*"

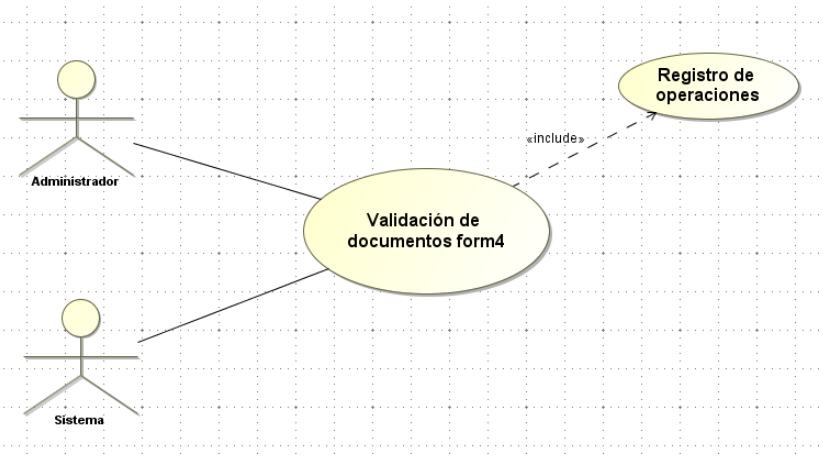


Ilustración 11 Caso de uso Validación de documentos "form 4"

CAU6	Registro de operaciones

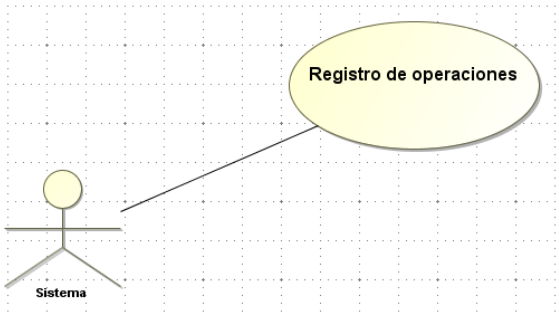


Ilustración 12 Caso de uso Registro de operaciones

CAU7	Alertas sobre eventos

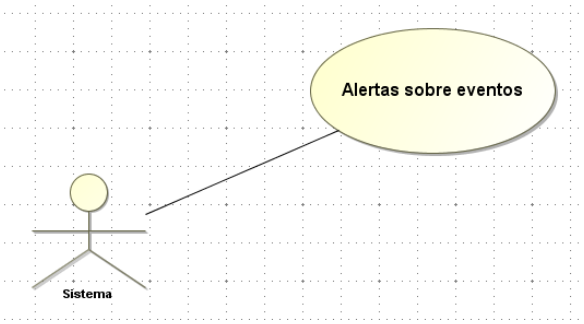


Ilustración 13 Caso de uso Alertas sobre eventos

CAU8	Análisis de datos

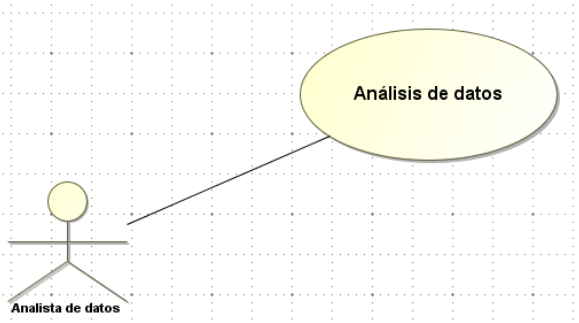


Ilustración 14 Caso de uso Análisis de datos

3 DISEÑO DEL SISTEMA

Al ser un proyecto que está basado en el desarrollo de un *"data warehouse",* es necesario definir un modelo de dominio que pueda representar de forma completa toda la información contenida en un documento *"form 4"*, para posteriormente, poder extraer información analítica de todo el conjunto de documentos almacenados.

El diseño conceptual, expresa en forma de modelo de dominio, las clases y sus distintas relaciones que posteriormente se trasladaran a un diseño lógico y físico para poder desarrollar una base de datos relacional que permita almacenar toda la información.

Como veremos en este apartado, la información que proporciona un documento *"form 4"* es muy extensa y aunque para el desarrollo del proyecto no es necesaria en su totalidad, se ha querido incluir toda la información con el objetivo de construir un *"data warehouse"* que pueda cubrir futuras necesidades de análisis, como por ejemplo, poder examinar las ejecuciones de "*stock options"* meses antes de su expiración[11] y así poder advertir comportamientos anómalos que puedan ser significativos.

A su vez, también se define el modelo de dominio del software "ETL" que se encargará de realizar las operaciones de extracción, transformación y carga de datos en el *"data warehouse"*.

3.1 DISEÑO CONCEPTUAL

3.1.1 *"Data WAREHOUSE"*

El diseño conceptual del *"data warehouse"* se ha dividido en las cuatro partes que conforman un documento "*form 4*" y una parte que almacena información analítica:

- **"Filing":** Información general del documento. Contiene información del tipo de documento, la fecha de presentación o las notas adicionales.

- **"Owner":** Información del "*insider"* que publica la información. Contiene información sobre el código único de identificación (**CIK**) dentro del sistema "EDGAR", así como el nombre, cargo y descripción del rol en la empresa.

- **"Issuer":** Información de la compañía en la que existe una relación con el "*insider".* Contiene información sobre el código único de identificación (**CIK**) dentro del sistema "EDGAR" y su símbolo de cotización en el mercado bursátil.

[11] Gabele, B, « When Insiders Exercise Options, Look Out»,1999, https://www.thestreet.com/opinion/when-insiders-exercise-options-look-out-756167

- **“Transaction”:** Información sobre la transacción reportada. En el caso de las transacciones de compra o venta, contiene datos como el total de acciones, el precio o el tipo de propiedad de las acciones.

- **“Transaction Result”:** Información sobre el resultado de la operación en diferentes periodos de tiempo.

Para desarrollar el diseño conceptual se realizado un trabajo de análisis sobre multitud de documentos *“form 4”* para poder extraer un modelo de dominio que pueda cubrir todos los casos que puedan producirse.

3.1.1.1 “FILING”

En la ilustración 15 podemos ver la estructura de un documento *“form 4* XML” y los metadatos asociados que se adjuntan a este documento para su clasificación en el sistema “EDGAR”.

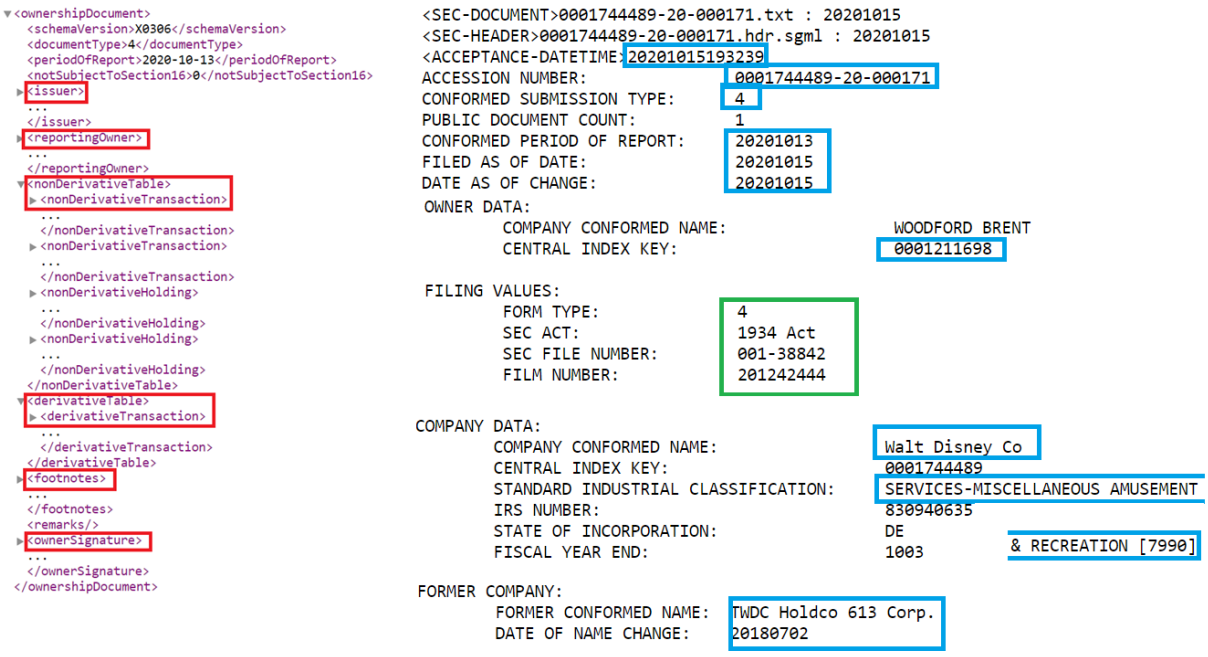


Ilustración 15 Estructura base de un documento “form 4 XML” (izquierda) y metadatos “EDGAR” del envío (derecha)

Existen 2 tipos de documentos *“form 4”*: el tipo **4** y el tipo **4/A**. El 4/A se definió para proporcionar a las empresas un mecanismo de corrección de errores en documentos presentados previamente, añadiendo un campo adicional que menciona la fecha en la que se presentó el documento original incorrecto y definiéndolo como tipo 4/A.

En la representación del modelo de dominio de la ilustración 16 se han definido todos los elementos básicos que deben almacenarse de un documento *“form 4”*. En el

"data warehouse" se almacenarán detalles sobre la fecha de presentación del documento, el código único ("*accession number"*) en el sistema "EDGAR" para evitar documentos duplicados, así como notas generales que se adjuntan al documento "***Remarks***" y que proporcionan información adicional sobre el envío.

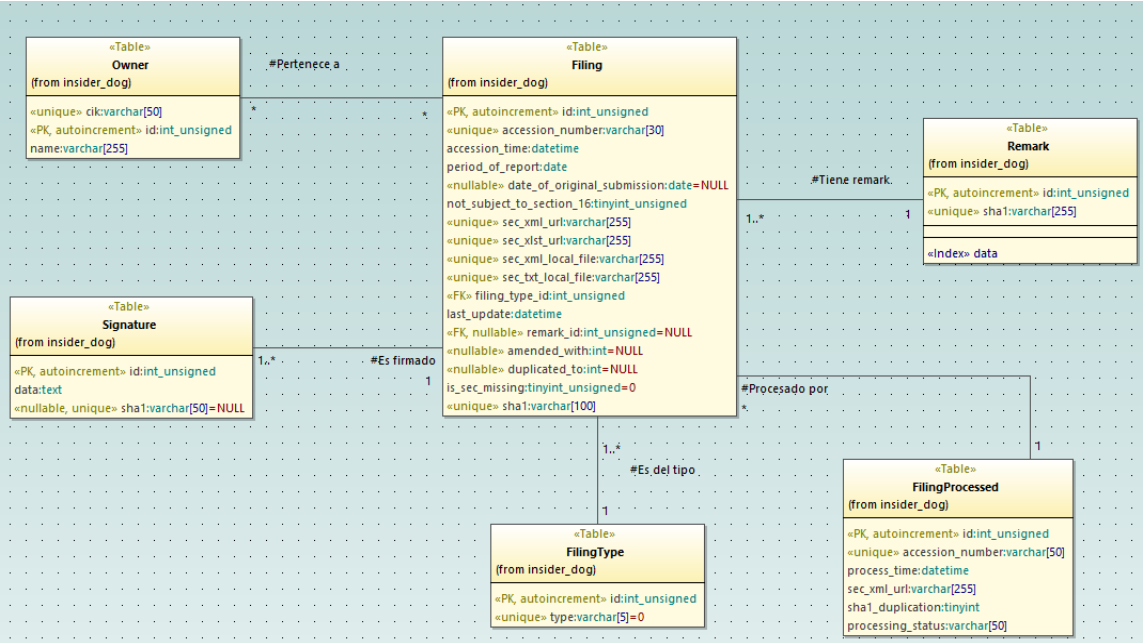


Ilustración 16 Diseño conceptual de un "Filing"

3.1.1.2 "OWNER"

La información más importante del "*insider"*, denominado "*Owner"* en el documento "*form 4*", es su identificación única dentro del sistema "EDGAR" y el rol que desempeña dentro de la compañía (director, accionista, etc.) así como una descripción detallada del rol, por ejemplo, si es el director financiero "*Chief Financial Officer"* y tal como veremos en la fase de análisis, es un dato muy importante para poder asignar fiabilidad a la información presentada. Como vemos en la ilustración 17, al no estar esta información normalizada, será necesario realizar operaciones de limpieza para que se pueda extraer un valor analítico.

```
<reportingOwner>
  <reportingOwnerId>
    <rptOwnerCik>0001726322</rptOwnerCik>
    <rptOwnerName>Vosseller Leigh</rptOwnerName>
  </reportingOwnerId>
  <reportingOwnerAddress>
    <rptOwnerStreet1>C/O TANDEM DIABETES CARE, INC.</rptOwnerStreet1>
    <rptOwnerStreet2>11075 ROSELLE STREET</rptOwnerStreet2>
    <rptOwnerCity>SAN DIEGO</rptOwnerCity>
    <rptOwnerState>CA</rptOwnerState>
    <rptOwnerZipCode>92121</rptOwnerZipCode>
    <rptOwnerStateDescription/>
  </reportingOwnerAddress>
  <reportingOwnerRelationship>
    <isDirector>0</isDirector>
    <isOfficer>1</isOfficer>
    <isTenPercentOwner>0</isTenPercentOwner>
    <isOther>0</isOther>
    <officerTitle>EVP & CHIEF FINANCIAL OFFICER</officerTitle>
    <otherText/>
  </reportingOwnerRelationship>
</reportingOwner>
```

Ilustración 17 Datos sobre el insider en el documento "form 4 XML"

También es interesante mencionar que un "*insider*" puede tener diferentes roles dentro de una empresa, por ejemplo, puede ser vicepresidente ejecutivo y a su vez ser el director financiero, así como pertenecer a diferentes empresas diferentes y no tiene por qué ser una persona física, sino que puede ser una persona jurídica, en forma de sociedad de inversión u otro tipo de entidad.

En la ilustración 18 podemos ver el diseño conceptual para almacenar toda la información comentada y sus relaciones.

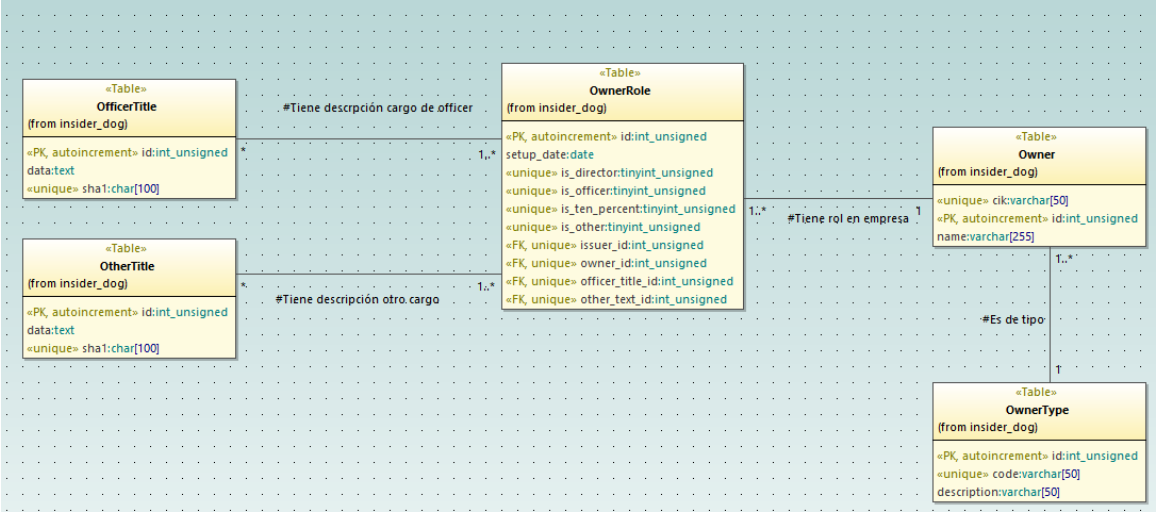


Ilustración 18 Diseño conceptual de un "insider" ("Owner")

3.1.1.3 "ISSUER"

Los datos de la empresa que reporta la información al sistema "EDGAR", entidad denominada "*issuer*" en el documento "*form 4*", no son muy extensos pero sí muy importantes ya que sirven para la identificación y clasificación de una compañía dentro del sistema.

De la misma forma que al procesar la información de los *"insiders"*, en muchos casos, será necesario realizar una limpieza de los datos previa a su ingesta por el sistema ya que aunque el "XML" obliga a las empresas a comunicar el símbolo de cotización en el mercado, no se controla si se ha informado correctamente, siendo necesario comprobar que el símbolo existe y en caso negativo, localizar el símbolo correcto y corregirlo previamente a su almacenamiento dentro del *"data warehouse"* tal como podemos ver en la ilustración 19.

```
▼<issuer>
   <issuerCik>0001744489</issuerCik>
   <issuerName>Walt Disney Co</issuerName>
   <issuerTradingSymbol>DIS</issuerTradingSymbol>
 </issuer>
```

Ilustración 19 Datos sobre el "issuer" en el documento "form 4 XML"

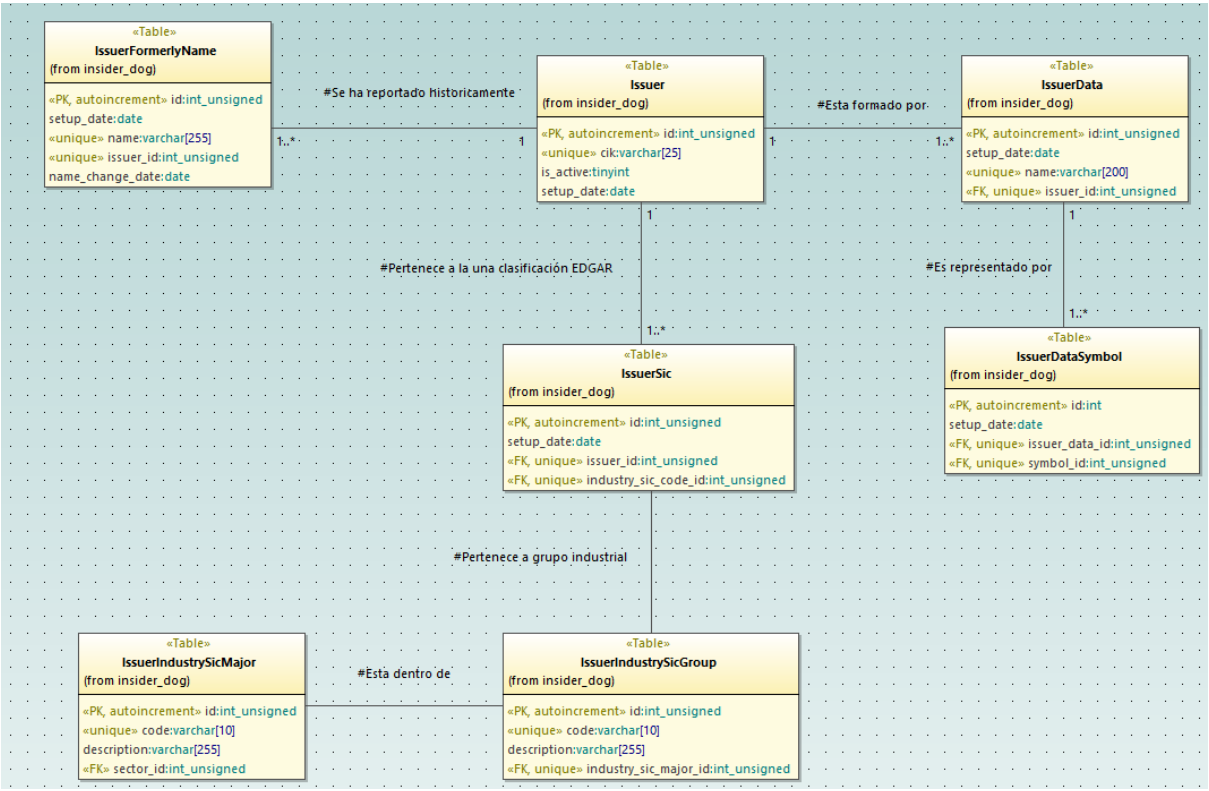


Ilustración 20 Diseño conceptual de una empresa ("Issuer")

En la ilustración 20 podemos ver el diseño conceptual para almacenar la información de la empresa y el sector industrial al que pertenece, información que no está contenida en el documento *"form 4"* pero sí que está presente en los metadatos "EDGAR" del envío.

También se refleja la posibilidad que una empresa, a lo largo de su vida, pueda tener distintos nombres y símbolos, todo ello siempre bajo un código único de identificación (CIK) en el sistema "EDGAR".

3.1.1.4 "TRANSACTION"

De toda la información que proporciona un documento *"form 4"*, el elemento más importante para el proyecto es el análisis de la transacción o transacciones de compra que se reporten, ya que esta información analizada en su conjunto nos proporcionará la información para poder predecir el futuro de una acción o índice bursátil.

El sistema "EDGAR", a través del formato "XML", nos proporciona 2 tipos de transacciones: derivativas y no derivativas. Las transacciones **no derivativas** "***NonDerivativeTransaction***" son en las que están incluidas todas las operaciones de compra o venta de acciones. Las transacciones **derivativas** "***DerivativeTransaction***"

son todas aquellas que reflejan operaciones con instrumentos financieros derivados, como por ejemplo, las opciones sobre acciones "***stock options***".

Aunque para el desarrollo del proyecto únicamente **son necesarias las operaciones de compra y por lo tanto las transacciones no derivativas**, se ha decidido almacenar todas las transacciones, incluidas las derivativas, con el futuro objetivo de poder estudiarlas utilizando modelos de análisis predictivo basados en "*machine learning*".

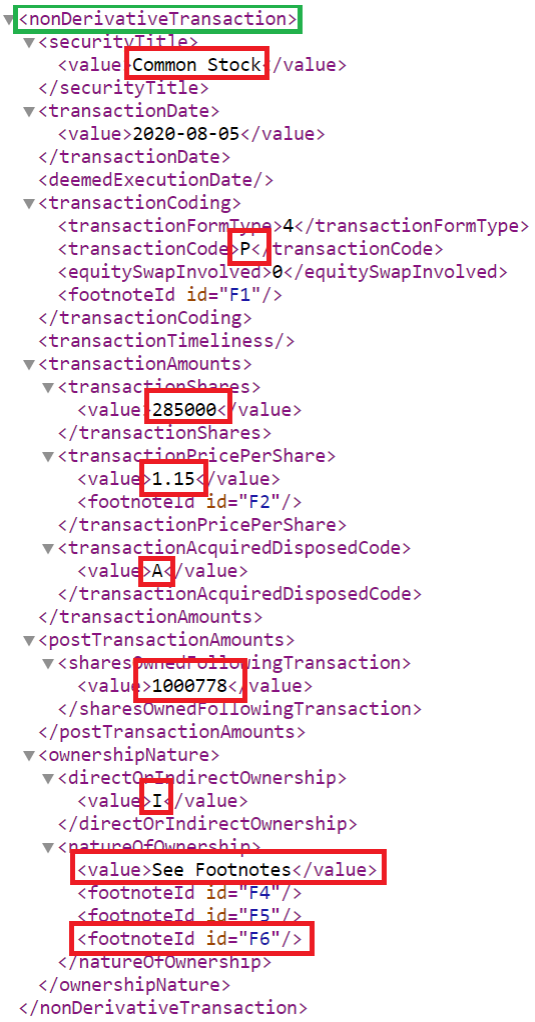


Ilustración 21 Datos sobre una transacción de compra (P) en el documento "form 4 XML"

Como podemos ver en la ilustración 21, de las transacciones, la información más importante es la fecha en la que se produjo la transacción, el tipo de transacción (**P** en el caso de compra), el total de acciones, el precio de compra, el tipo de propiedad, es decir si es comprada de forma directa o indirecta (a través de personas relacionadas con el *insider*) y las notas adicionales que se han especificado en el "tag" "***FootNotes***" y que pueden aportar claridad sobre la transacción que se está reportando, y descartar todas aquellas transacciones de compra relacionadas con operaciones sobre los planes sobre opciones de compra "*stock options*" que no son relevantes.

En la ilustración 22 se muestra el diseño conceptual a través de las clases y sus relaciones para poder almacenar toda la información en el *"data warehouse"*.

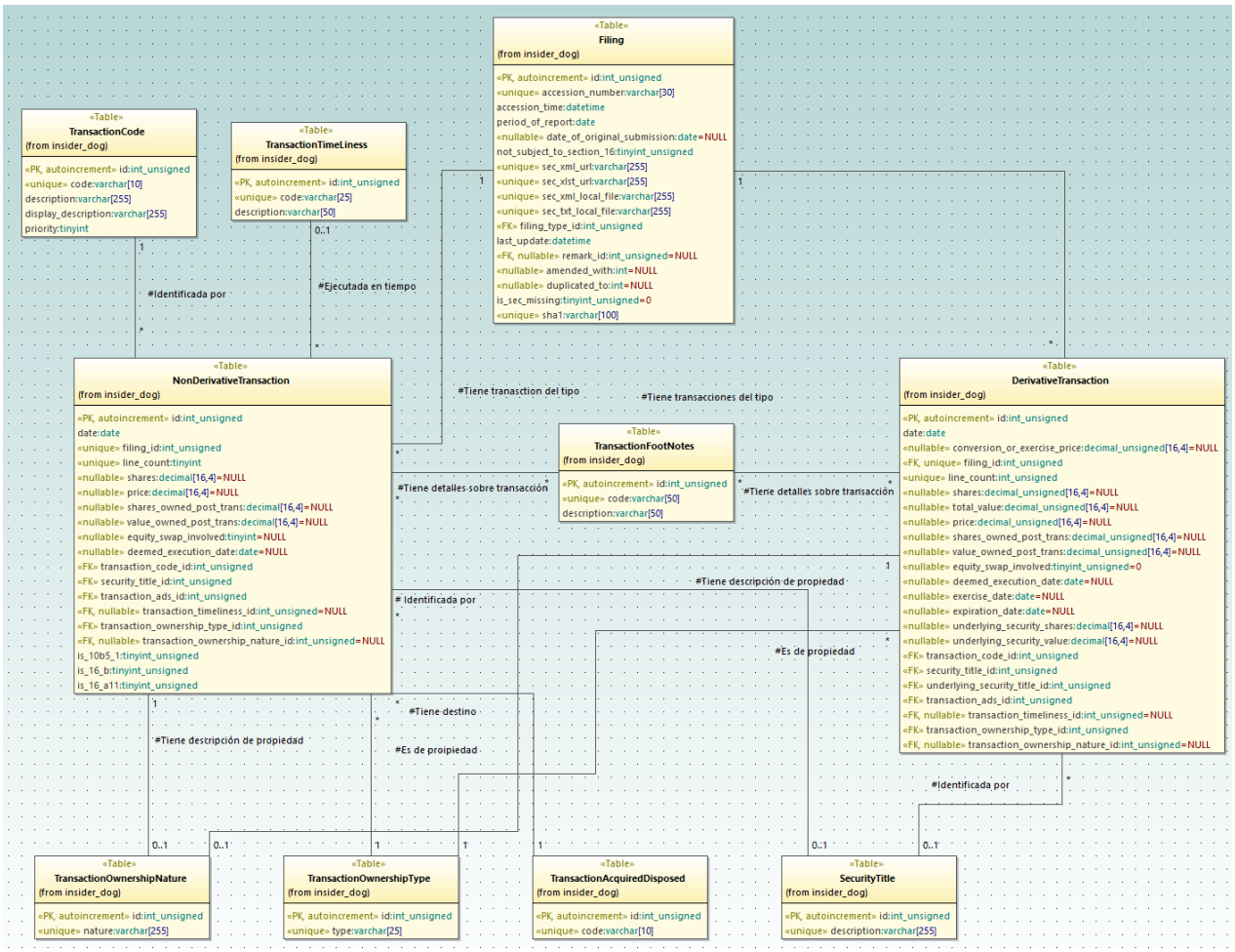


*Ilustración 22 Diseño conceptual de las transacciones **derivativas** ("DerivativeTransaction") y **no derivativas** ("NonDerivativeTransaction")*

3.1.1.5 "TRANSACTION RESULT"

Cada transacción de compra, tendrá asociado un resultado de operación. Este resultado, refleja el éxito o fracaso de la transacción desde el punto de vista del inversor, es decir, si un "*insider"* ha comprado 1000 acciones a 40 dólares el 10-10-2020 y el 10-10-2021 las acciones de la empresa cotizan a 80 dólares, el resultado de la operación es un éxito desde el punto de vista del inversor, ya que, replicando el comportamiento del "*insider",* el inversor podría haber obtenido un beneficio.

Esta información, que debe almacenarse en el *"data warehouse"* y que permite analizar las operaciones de forma individual, categorizando a cada "*insider"* basándose en el resultado de sus decisiones de inversión, deberá realizarse en futuras iteraciones

del proyecto ya que al ser un proyecto académico no se dispone del tiempo ni de los recursos económicos necesarios para poder almacenar estos resultados dada la necesidad de disponer de un sistema automatizado de cotizaciones del mercado bursátil.

En la ilustración 23 puede observarse el modelo de dominio para poder almacenar estos resultados de las transacciones en el *"data warehouse"*.

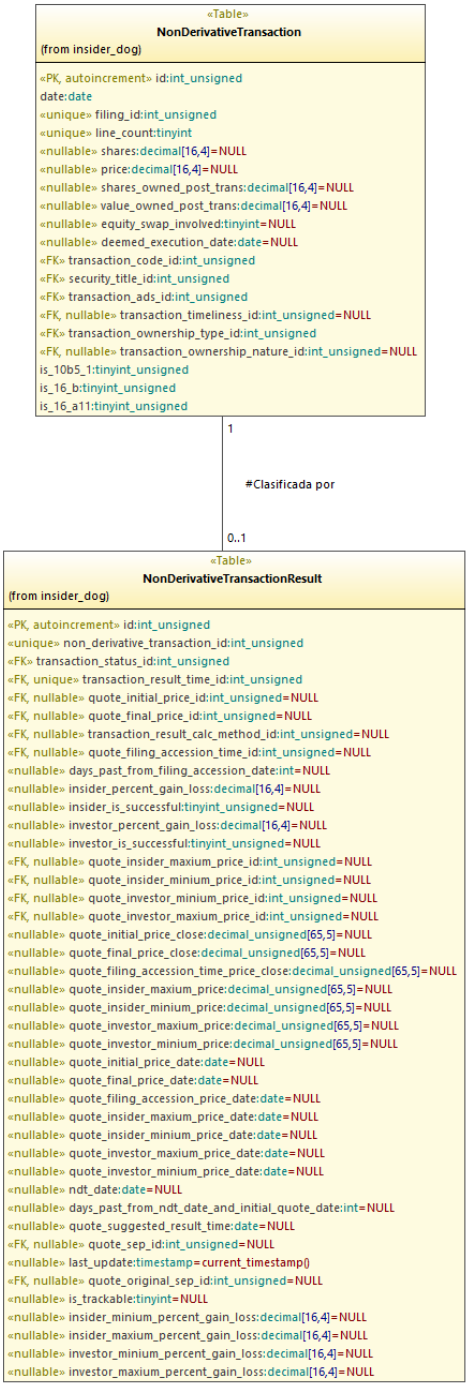


Ilustración 23 Diseño conceptual para almacenar los resultados de las transacciones de compra o venta

3.1.2 SOFTWARE "ETL"

3.1.2.1 OPERACIONES DE DESCARGA DE DOCUMENTOS *"FORM 4"*

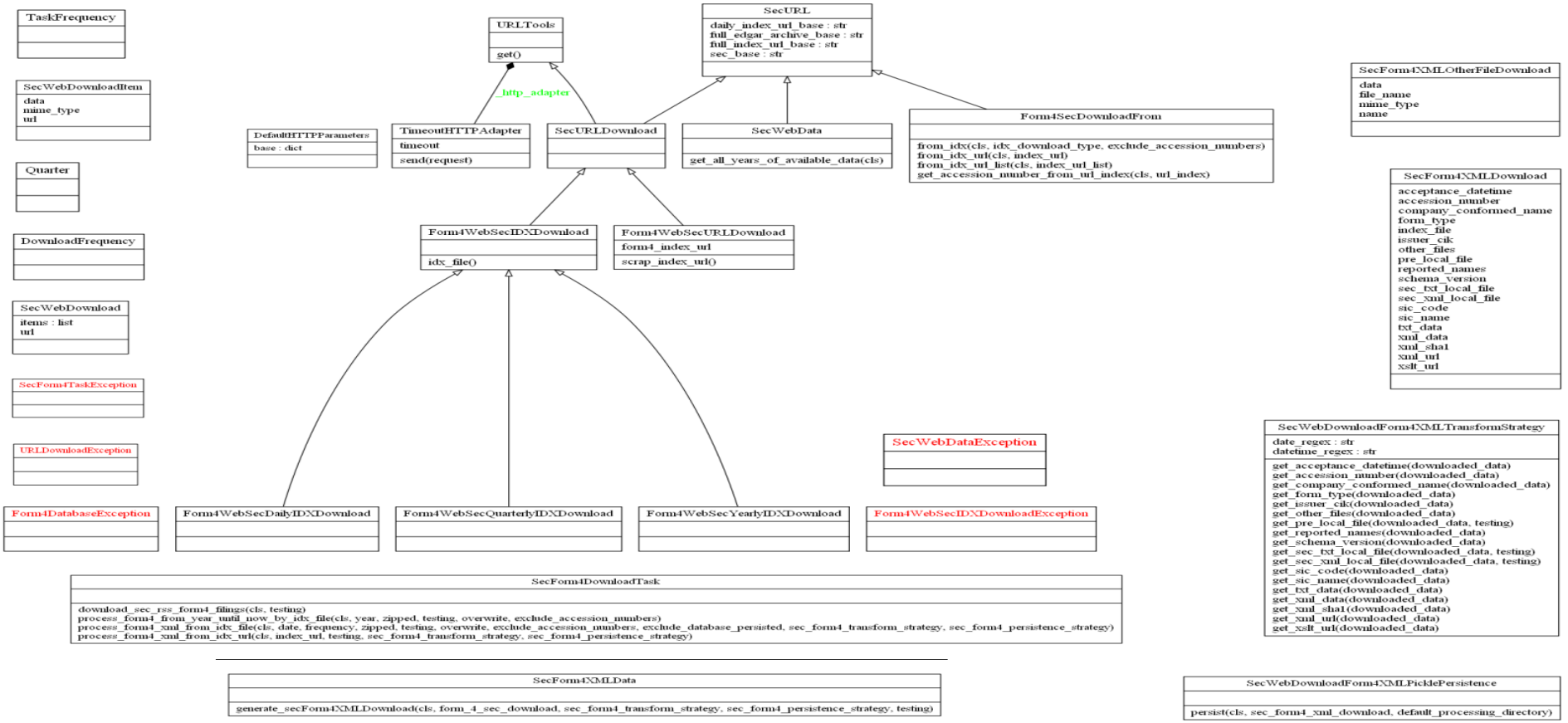


Ilustración 24 Modelo de dominio del software "ETL" en las operaciones de descarga de documentos "form 4" del sistema "EDGAR"

El software que alimentará el "*data warehouse*" se implementará como un software que se ejecute a través de la línea de comandos y por lo tanto se podrán ejecutar diferentes acciones como la descarga de documentos *"form 4"* o las operaciones "ETL" pasando diferentes parámetros.

En la ilustración 24 se muestra todo el modelo de dominio de las operaciones de descarga de un documento *"form 4"*.

Como se ha comentado anteriormente, para descargar un documento "*form 4*" del sistema "EDGAR", se pueden utilizar dos opciones, la descarga con el sistema "RSS" que es en tiempo real y la descarga con el servicio "IDX" que permite la descarga al final de un día laboral de todos los *"form 4"* publicados. La descarga "IDX", requiere una implementación más costosa, pero permite mayor flexibilidad, ya que podemos pedir al servicio "IDX" descargar documentos de todo un año completo.

Para el proceso de descarga se ha definido una clase base, que representa una descarga dentro del sistema y en la que se incluyen los metadatos del sistema "EDGAR" que el *"data warehouse"* necesita. Los más importantes son los siguientes:

- ***"Acceptance datetime"***: Fecha en la que se aceptó el documento en el sistema "EDGAR".
- ***"Accession number":*** Número de identificación único de un documento *"form 4"* en el sistema.
- ***"Schema version":*** Versión del esquema "XSD" en el que se ha transmitido la información. El esquema "XSD" en el sistema "EDGAR" ha evolucionado con los años y podemos encontrar variaciones en los documentos "XML" dependiendo del esquema utilizado.
- ***"Form Type":*** Tipo de documento *"form 4"*. Como se ha comentado, existe el 4 y el 4/A y es necesario diferenciarlos para poder validar correctamente la información del "XML" ya que cada tipo tiene su propio *schema*.
- ***"XML data":*** Datos "XML" del documento, que incluyen toda la información sobre el documento *"form 4"*.
- ***"TXT data":*** Metadatos en formato de texto plano, que se guardan como garantía de la recepción.

Hay que recordar que un *"data warehouse"* no sólo ha de contener la información que se necesita en un momento determinado para un análisis, sino que su principal objetivo ha de ser el poder extraer información para futuros análisis, por lo tanto, su definición ha de ser lo más completa posible, tal y como se ha diseñado en este proyecto.

Se ha utilizado para su definición el uso del patrón "***Strategy***", lo que permite intercambiar diferentes algoritmos para poder obtener los datos necesarios para que un documento descargado sea válido, así como diferentes estrategias de persistencia, cumpliendo los principios de reusabilidad y extensibilidad. Esto permite, que en tiempo de ejecución, se pueda definir la estrategia de descarga que se aplicará, ya sea "RSS" o "IDX" y que el resultado de este cambio sea siempre el mismo. Si en el futuro, el sistema "EDGAR" cambia la forma de entregar la información, como por ejemplo, en formato "JSON", simplemente habrá que definir una nueva estrategia de extracción y el sistema no deberá ser modificado.

Cada documento descargado, por defecto, se almacenará en disco en un formato especial que será posteriormente importado por la opción "ETL" del software.

3.1.2.2 OPERACIONES DE EXTRACCIÓN Y TRANSFORMACIÓN DE DOCUMENTOS *"FORM 4"*

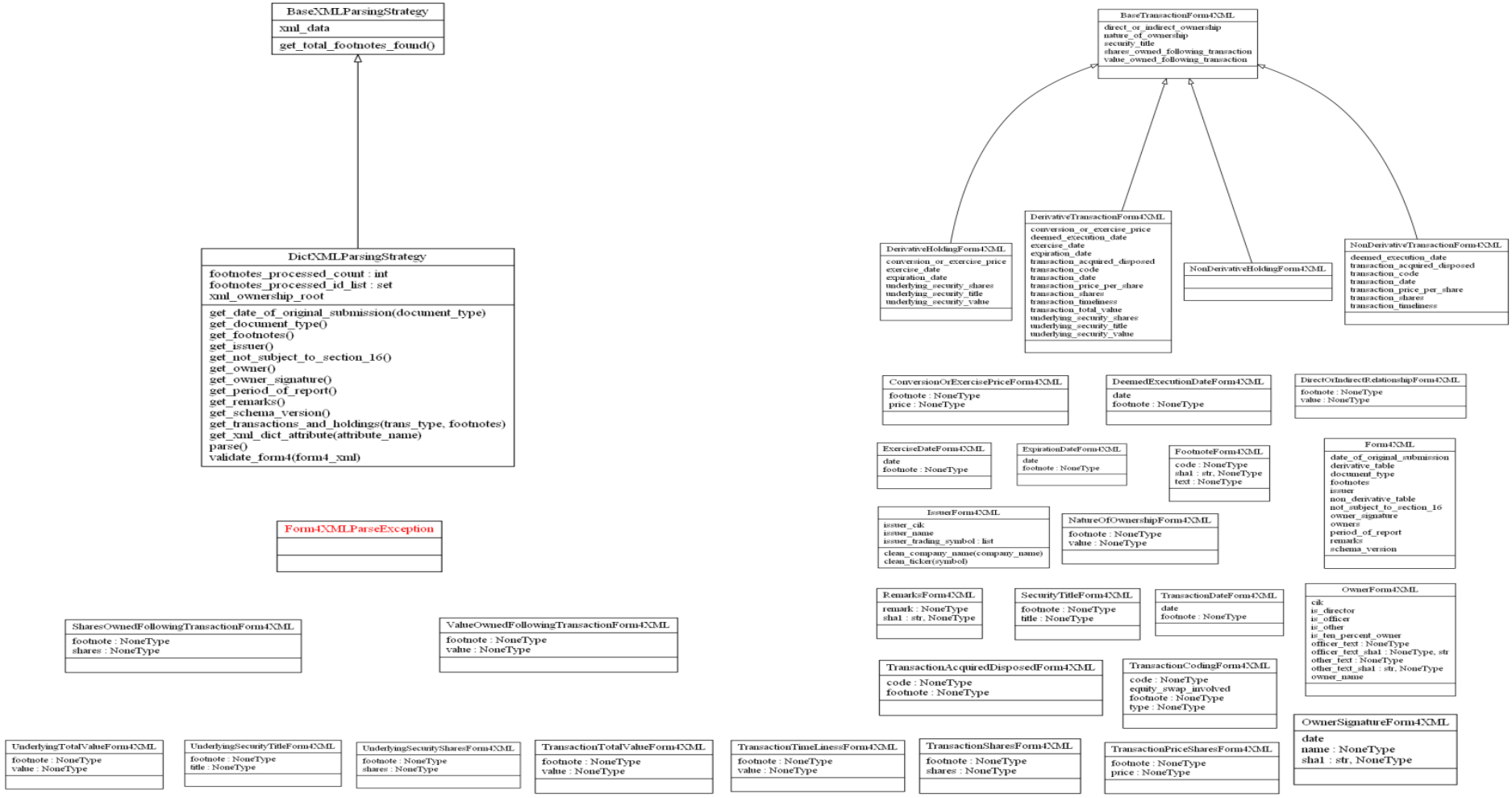


Ilustración 25 Modelo de dominio del software "ETL" en las operaciones de extracción y transformación de documentos "form 4" del sistema "EDGAR"

En la ilustración 25 se muestra el modelo de dominio del software "ETL", concretamente la operación de extracción y transformación de los documentos *"form 4"* en formato "XML".

De la misma forma que en el modelo de dominio para la descarga de documentos, el diseño aplica el uso del patrón "***Strategy***".

Para poder extraer información de un documento en formato "XML", podemos utilizar diferentes técnicas que pueden ser implementadas a través de diferentes algoritmos de extracción. En el modelo de dominio, se utilizará un único algoritmo basado en una estrategia que utiliza un mapa contenedor para la extracción de los datos.

Se ha definido un conjunto de clases que representan el modelo de dominio de un documento *"form 4"* representado por los datos "XML" y que a través de sus operaciones, se realizarán las diferentes tareas de transformación de los datos, proporcionando al diseño el principio de responsabilidad única.

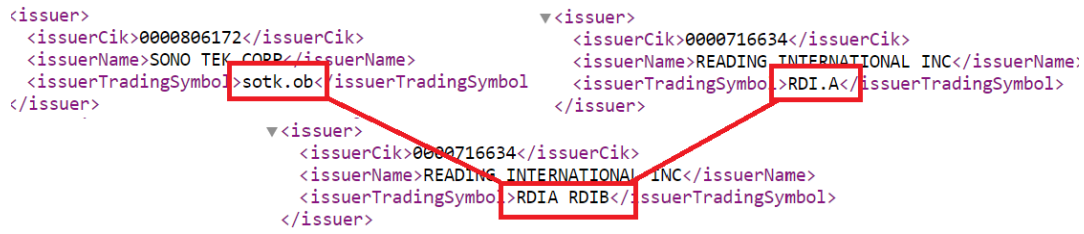


Ilustración 26 Ejemplo de Transformación necesaria para lo símbolos de cotización

Como podemos ver en la ilustración 26, cuando se proceda a la extracción de los símbolos que utiliza la empresa para su cotización en el mercado bursátil, es muy posible que el símbolo sea incorrecto o que se ha informado más de uno, lo que implica la necesidad de una transformación previa para poder trabajar con ellos de manera relacional.

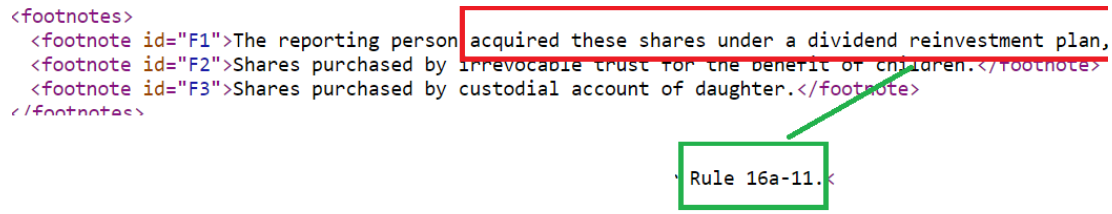


Ilustración 27 Información a transformar en identificador de transacción sin interés

Otro ejemplo de transformación se muestra en la ilustración 27 y es la identificación de un documento que ha sido remitido al sistema "EDGAR" advirtiendo, que la operación que se detalla, forma parte del plan de reinversión de dividendos[12] que ha realizado el "*insider*" y que debemos transformar en un identificativo que permita excluirlo de un posible análisis ya que son operaciones que no proporcionan ningún valor analítico.

[12] SEC.GOV, «R. SECTION 16 RULES AND FORMS 3, 4 AND 5 », https://www.sec.gov/interps/telephone/cftelinterps_sec16.pdf

3.1.2.3 OPERACIONES DE CARGA DE LA INFORMACIÓN EN UNA BASE DE DATOS RELACIONAL

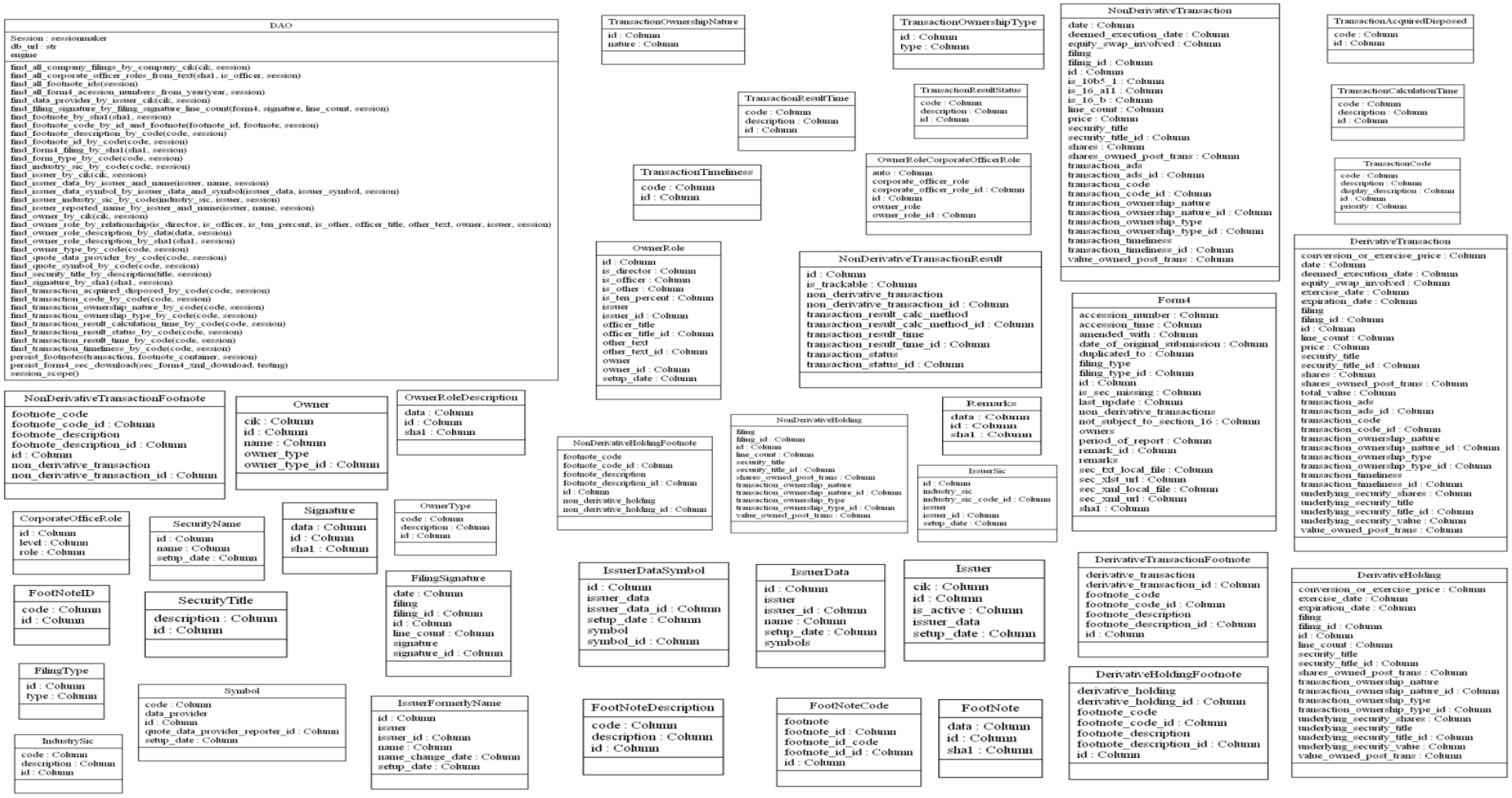


Ilustración 28 Modelo de dominio del software "ETL" en la operación de carga de documentos "form 4" del sistema "EDGAR"

En la ilustración 28 se muestra el diseño del modelo de dominio referente a las operaciones de carga de datos en una base de datos relacional.

Para el diseño se ha utilizado una clase base que aplica el uso del patrón "***Façade***" para estructurar el subsistema de persistencia.

Como se puede observar, el modelo de dominio de esta capa muestra todas las entidades que tienen representación en la base de datos, como por ejemplo, la clase "***OwnerRole***" que será la encargada de almacenar el rol de un "*insider"* dentro de la compañía reflejado en la información proporcionada por un documento "*form 4*".

3.2 DISEÑO FÍSICO DE LA BASE DE DATOS

El objetivo fundamental de este proyecto es el análisis de datos y la selección del modelo en el que se almacenaran los datos, es una decisión muy importante.

Existen muchas opciones para almacenar datos y para este proyecto se ha escogido el modelo relacional.

La elección de un modelo relacional garantiza, que si existe una buena definición en el diseño, a través de normalización, se podrán obtener, entre otras ventajas, resultados fiables en las operaciones de extracción de datos, eliminando uno de los principales problemas en el análisis de la información, la duplicación de datos.

El uso principal de un *"data warehouse"* es el almacenamiento de un histórico de datos para poder extraer información analítica y en el caso de este proyecto, es el sistema "EDGAR" el encargado de alimentar este histórico. Este sistema nunca actualiza la información proporcionada y si es necesario corregirla, se genera un nuevo documento *"form 4"*, del tipo 4/A por lo que no es necesario realizar ningún tipo de operación de actualización, únicamente será necesaria la persistencia de nuevos datos. Es por todo ello que en la fase de diseño se ha tenido en cuenta que el proyecto únicamente necesita obtener información analítica de datos históricos.

Una vez estudiado y definido el diseño lógico, basado en la información proporcionada por el diseño conceptual expuesto en el apartado 3.1.1, se ha se ha definido el diseño físico de la base datos en el que se puede apreciar como resultado del análisis la gran cantidad de tablas resultante.

Tal y como se ha comentado anteriormente, en el diseño, no se ha querido únicamente contemplar todo aquello estrictamente necesario para las conclusiones de este proyecto sino que se ha realizado un diseño que pueda ser un punto de partida para futuros análisis a nivel exploratorio o usando modelos de análisis predictivo. Es por este motivo que también se almacenará en el histórico la información de entidades que no tienen un papel fundamental para las conclusiones, como por ejemplo, las notas de los documentos *"form 4"* denominadas "***Footnotes***" o las transacciones que no son operaciones de compra.

En la ilustración 29, podemos ver el diseño físico del modelo relacional, así como todas las relaciones entre las tablas.

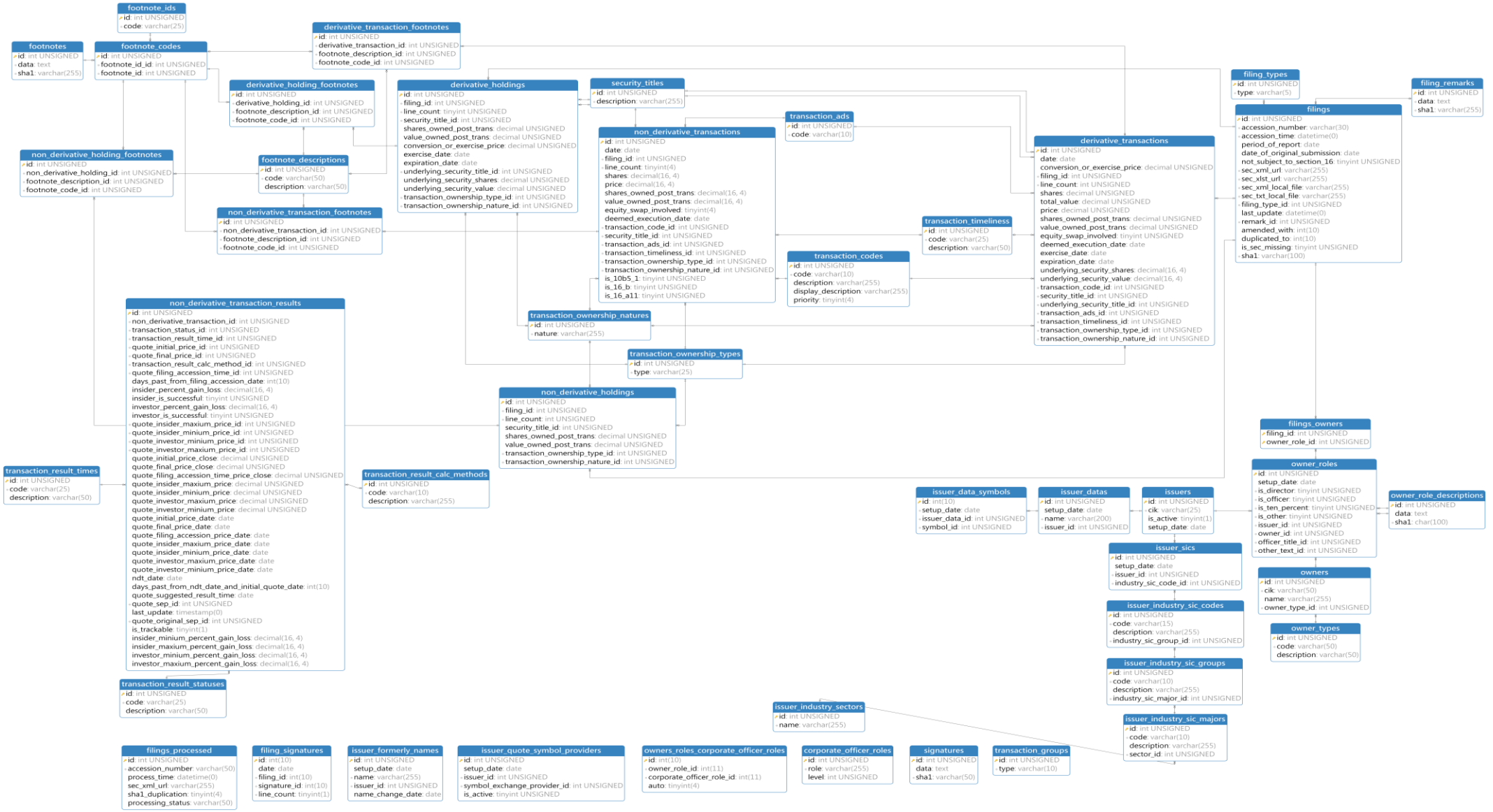


Ilustración 29 Diseño físico de la base de datos

4 IMPLEMENTACIÓN

4.1 INSTALACIÓN Y CONFIGURACIÓN DEL ENTORNO

El proyecto se basa fundamentalmente en el uso de 2 tecnologías "*Open source*": **Python y MariaDB**.

4.1.1 INSTALACIÓN DE PYTHON Y LAS LIBRERÍAS UTILIZADAS

El proyecto ha sido desarrollado utilizando la versión del lenguaje de programación **Python** en su versión 3.8.5 que es la versión estable más reciente.

De la misma forma que ocurre con otros lenguajes de programación, la instalación de las librerías que se usan en el proyecto debe realizarse a través de un entorno único, aislado del resto de proyectos.

Para simplificar la instalación y debido al uso intensivo que se va a realizar de la librería **Pandas** para hallar las conclusiones del análisis, se ha optado por definir una guía basada en la instalación de **Anaconda**, una distribución de Python, que incluye multitud de librerías para el análisis de datos. Utilizando esta distribución añadiremos alguna librería adicional que el proyecto necesita.

El desarrollo de esta guía se ha basado en la instalación en un sistema operativo Windows 10 de 64 bits.

4.1.1.1 INSTALACIÓN Y CONFIGURACIÓN DE LA DISTRIBUCIÓN ANACONDA

Como vemos en la ilustración 30, debemos descargar la distribución más reciente de Anaconda desde su sitio Web oficial: https://www.anaconda.com/products/individual y realizar su instalación.

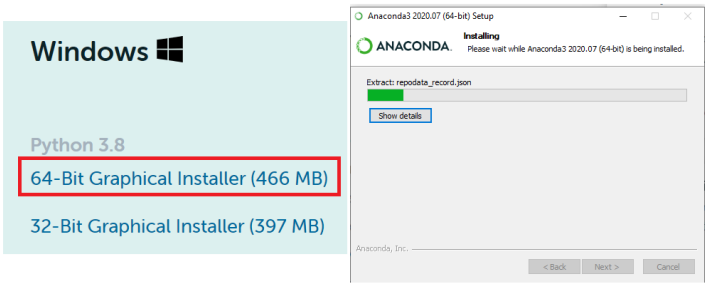
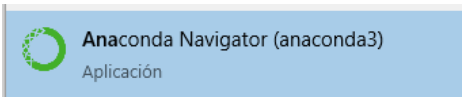


Ilustración 30 Descarga e Instalación de Anaconda

Una vez instalada la distribución, deberemos arrancarla haciendo ejecutando la aplicación "Anaconda Navigator":



Anaconda, utilizando herramientas muy conocidas por los desarrolladores Python, nos permite crear un entorno de ejecución del proyecto aislado, permitiendo instalar librerías y versiones específicas de estas.

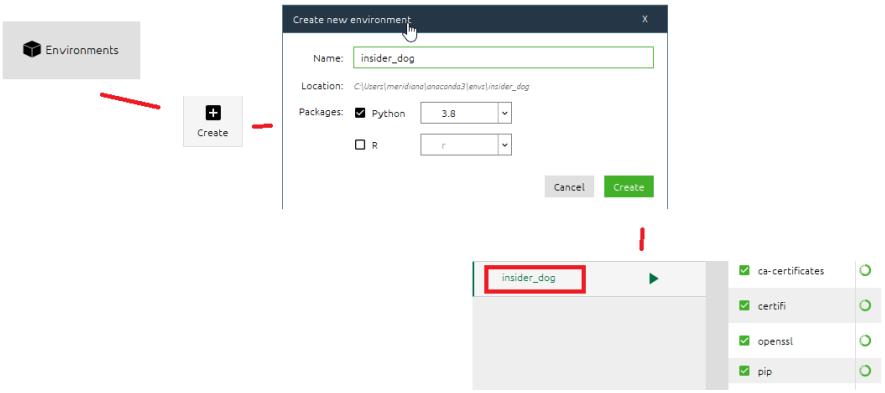


Ilustración 31 Creación de un entorno aislado de desarrollo y ejecución

4.1.1.2 INSTALACIÓN DE LAS LIBRERÍAS NECESARIAS

Como vemos en la ilustración 31, utilizando la opción "*Environments*" hemos creado un entorno aislado llamado "**insider_dog**" al que hemos establecido el uso por defecto de una versión concreta de Python y por defecto Anaconda nos ha instalado unas librerías base con las que empezar a trabajar.

El proyecto, necesita unas librerías adicionales que Anaconda no instala por defecto y que son las siguientes:

- SQLAlchemy
- "XML"todict
- "XML"schema
- urllib3
- requests
- PyMySQL
- l"XML"
- pyyaml
- feedparser
- click

Estas librerías, proporcionaran al software "ETL", capacidad de gestionar información en formato "XML" o gestionar la persistencia de los datos.

Para realizar la instalación de cada una de ellas en nuestro nuevo entorno creado, debemos ejecutar una consola de "*shell"* en Windows (**cmd**), usando la opción que nos proporciona Anaconda como podemos ver en la figura 32:

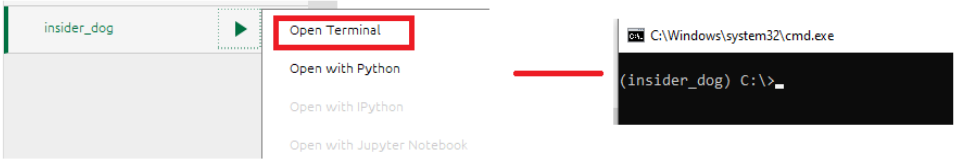


Ilustración 32 Ejecución de un terminal "shell" en Windows con el entorno de trabajo

Una vez en el "*shell"*, procedemos a instalar las librerías como se expone en la figura 33 necesarias para la ejecución del "ETL" utilizando el comando "***pip***" que incluye Python:

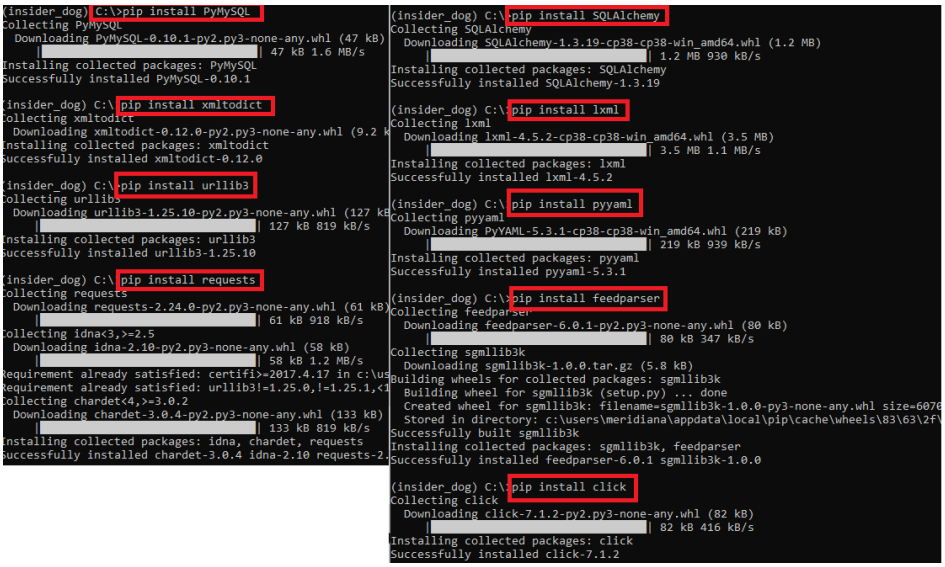


Ilustración 33 Instalación de las librerías Python necesarias usando el comando "pip"

Para el análisis exploratorio utilizaremos la librería Pandas, muy utilizada en el análisis de datos, que instalaremos tal y como podemos ver en la ilustración 34:

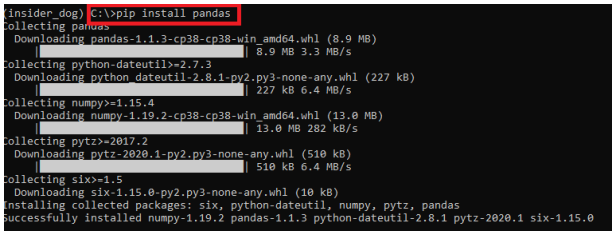


Ilustración 34 Instalación de la librería de análisis de datos Pandas

4.2 DESCARGA DEL HISTÓRICO DE DOCUMENTOS *"FORM 4"* (2003-2020)

Los documentos *"form 4",* deben descargarse previamente para ser procesados y cargados en el *"data warehouse"*. El software "ETL", a través de la línea de comandos, descargará los documentos en un directorio predeterminado, utilizando el formato de serialización de objetos Python, denominado "***pickle***". La extensión de un fichero descargado y no procesado será "**.pre"** que indica que el documento *"form 4"* está en un estado de preproceso.

Una vez el documento sea procesado y aceptado por el proceso "*data warehousing"*, se procederá a moverlo a un directorio dentro del disco, a modo de conservación histórica.

La ilustración 32 muestra el contenido de un fichero ".pre" una vez cargado, al que posteriormente se le aplicarán las operaciones "ETL".

Como se puede ver, el formato .pre contiene los datos del documento (datos "XML") así como todos los metadatos del sistema "EDGAR" necesarios para su posterior tratamiento.

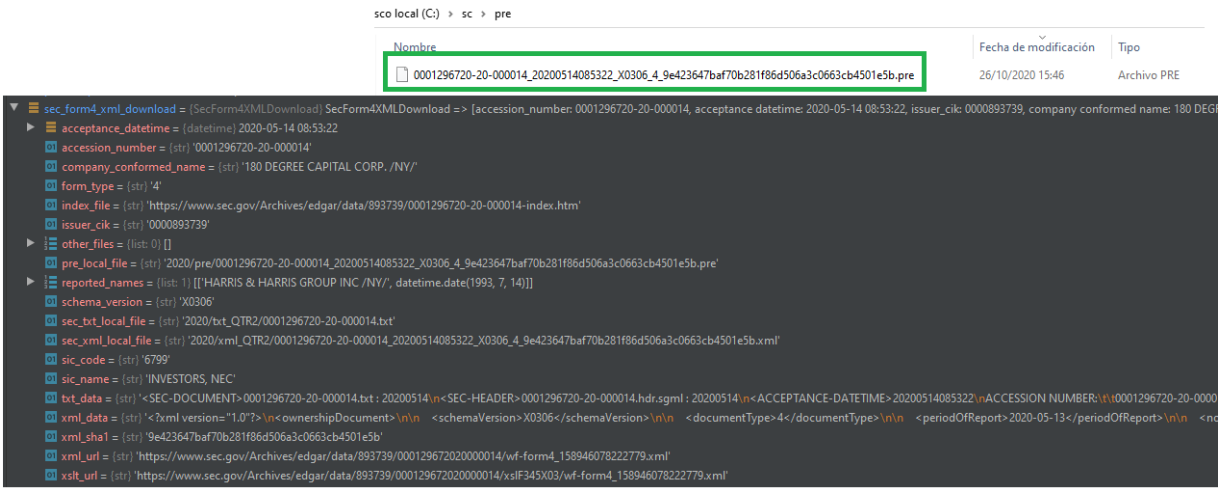


Ilustración 32 Contenido de un documento "form 4" ".pre" abierto para su proceso

4.2.1 INSTALACIÓN DEL SOFTWARE "ETL" Y PRUEBA DE DESCARGA

El código fuente del software "ETL" se puede encontrar en el proyecto con el nombre de **insider_dog.zip** y se deberá descomprimir en un directorio llamado "**d:\insider_dog**".

Para comprobar que el software "ETL" funciona correctamente, se va proceder a la descarga todos los documentos *"form 4"* de un día concreto. Es aconsejable descargar, como prueba, los del día anterior ya que no estarán importados en el sistema. Para ello, crearemos la carpeta destino "**c:\\sc\pre"** donde se almacenarán las descargas ".pre" (recordemos que el directorio destino de los archivos ".pre", se puede configurar dentro del software "ETL" y en concreto modificando la línea 109 del archivo **"utils.py"**)

Una vez creado el directorio y descomprimido el paquete que contiene el software “ETL”, vamos a lanzar el siguiente comando que nos permitirá descargar todos los form4 en formato .pre en el directorio “**c:\sc\pre**”:

```
python etl.py download-daily-form4-filings-from-idx-file --date 2020-05-14 --testing true --overwrite false
```

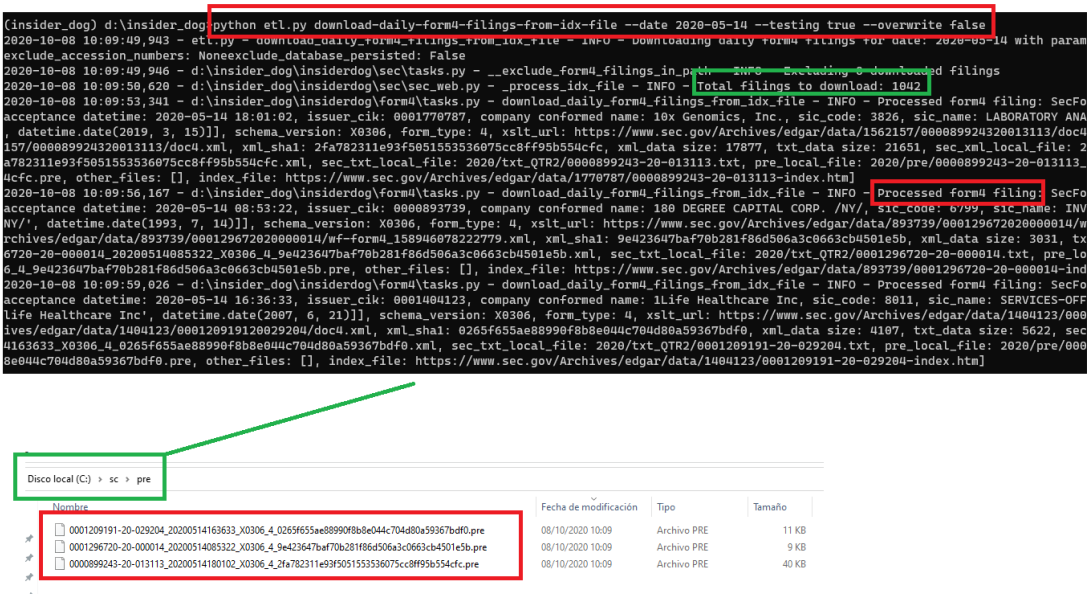


Ilustración 35 Descarga de documentos “form 4” de un día concreto

En la ilustración 35, podemos comprobar que la ejecución del software “ETL” con los parámetros especificados ha descargado un total de **1042** documentos “.pre” en el directorio que previamente habíamos establecido como directorio de descarga.

Dado que el software “ETL” basa su ejecución en diferentes comandos, en la siguiente sección procederemos a descargar el histórico y posteriormente, una vez hayamos instalado la base de datos, podremos comprobar el funcionamiento completo de las operaciones “ETL”, ejecutando los comandos que realizan la extracción, transformación y carga de estos documentos descargados en la base de datos relacional.

4.2.2 DESCARGA DEL HISTÓRICO UTILIZANDO EL SERVICIO “IDX”

Como hemos definido en la fase de diseño, una de las operaciones que proporciona el software “ETL” es la posibilidad de descarga de los documentos ya sea a través del servicio “RSS” o “IDX”.

Para descargar el histórico de documentos, comprendido entre los años 2003 hasta el 2020, utilizaremos el servicio “IDX” del sistema “EDGAR”, al que se especificará una fecha concreta y el software “ETL” descargará todos los documentos *“form 4”* que incluya el índice para ese año.

La descarga de documentos a través del software “ETL” tiene tres opciones de uso principales:

- **Descarga de documentos en forma diaria (Servicio “IDX”):** Se proporciona una fecha concreta y el software descarga los documentos en esa fecha.

- **Descarga de documentos en un año concreto (Servicio “IDX”):** Se proporciona una fecha concreta y el software extrae el año de esa fecha procediendo a la descarga todos los documentos de todo el año.

- **Descarga de documentos en tiempo real (Servicio “RSS”):** Al ejecutar el comando con el parámetro “RSS” el software descargará todos aquellos documentos que están pendientes de introducir en el sistema y que acaban de ser publicados en el sistema “EDGAR”.

Para la descarga de documentos utilizando el software se debe ejecutar el comando **python etl.py** utilizando los siguientes parámetros base:

- **download-daily-form4-filings-from-idx-file:** Operación de descarga de documentos *“form 4”* **diarios** (parámetro base)

- **download-yearly-form4-filings-from-idx-file:** Operación de descarga de documentos *“form 4”* **anuales** (parámetro base)

- **download_sec_rss_form4_filings”:** Operación de descarga de documentos *“form 4”* **en tiempo real** (parámetro base)

La ejecución de todos estos parámetros base acepta los siguientes parámetros:

- **--date:** Parámetro que acepta un argumento del tipo fecha que se utilizará para extraer el año para la descarga de los documentos *“form 4”*. En el caso del servicio “RSS”, el parámetro es ignorado.

- **--testing:** Parámetro que acepta un argumento del tipo (true/false) y que indica si se está ejecutando en un entorno de pruebas.

- **--overwrite:** Parámetro que acepta un argumento del tipo (true/false) y que indica si se han de sobrescribir archivos que ya se habían descargado. El software se ha diseñado para ser capaz de retomar la descarga en un punto, sin tener que volver a descargar todos los documentos, si el proceso se interrumpe.

- **--exclude_database_persisted:** Parámetro que acepta un argumento del tipo (true/false) y que excluye de la descarga todos aquellos documentos que ya están en el “*data warehouse”*. Es un parámetro básico para descargar todos aquellos documentos nuevos en el sistema “EDGAR”.

A continuación vemos un ejemplo del resultado de la descarga de los datos en el periodo 2003-2020:

2003:

La descarga de documentos *"form 4"* del año 2003 merece una especial atención ya que no fue hasta el 05-05-2003 cuando se empezó a distribuir la información en formato "XML" a través del sistema "EDGAR", por lo tanto, el año 2003 tiene un número inferior de documentos en formato "XML" al resto de todo el periodo. Por ello es necesario ejecutar el proceso de descarga diario desde el 05-05-2003 hasta el 31-12-2003. El sistema "EDGAR" no proporciona ninguna información para indicar si una fecha contiene información "XML" y por lo tanto la única opción es la descarga diaria para el año 2003. El resto del periodo se realizará con la opción de descarga anual, que como se puede ver en las ilustraciones se especifica la fecha de la descarga, el número de documentos del sistema "EDGAR" y un fragmento de los ficheros ".pre" descargados.

```
python etl.py download-daily-form4-filings-from-idx-file --date 2003-05-14 --testing true --overwrite false

python etl.py download-daily-form4-filings-from-idx-file --date 2003-05-15 --testing true --overwrite false

……
```

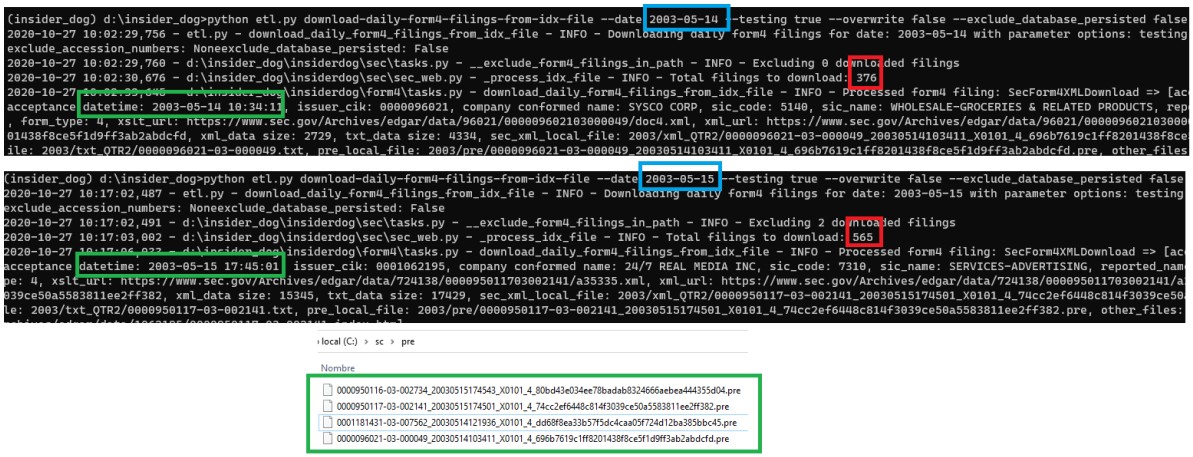


Ilustración 33 Descarga diaria de documentos "form 4" para el año 2003

2004:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2004-10-14 --testing true --overwrite false
```

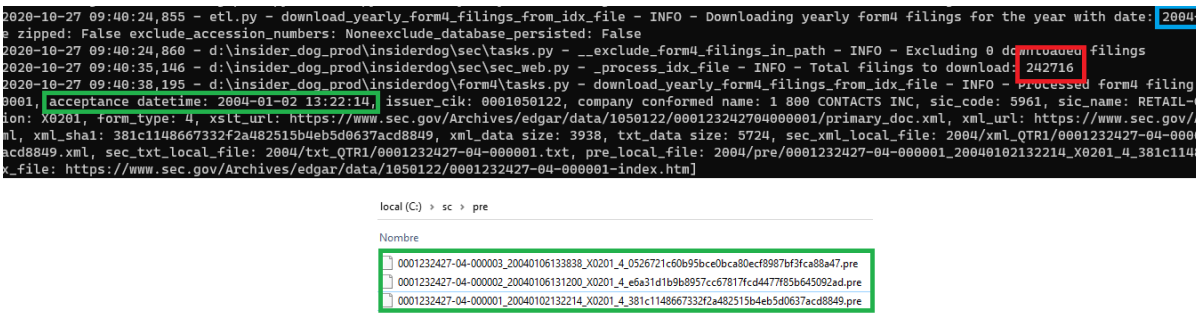


Ilustración 34 Descarga documentos "form 4" para el año 2004

2005:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2005-10-14 --testing true --overwrite false
```

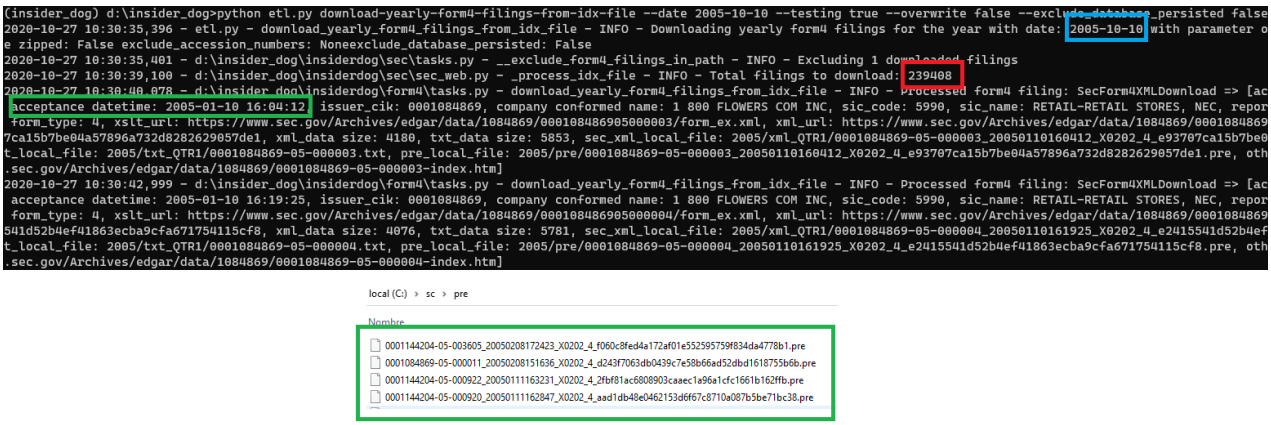


Ilustración 35 Descarga documentos "form 4" para el año 2005

2006:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2006-10-14 --testing true --overwrite false
```

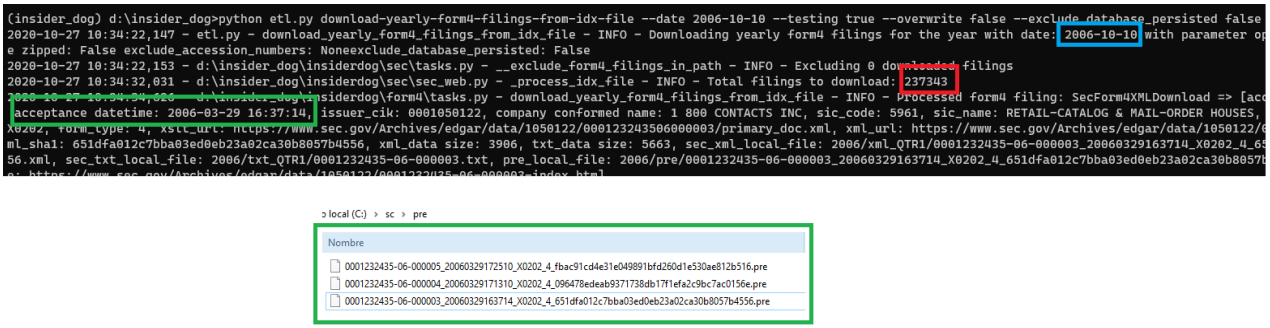


Ilustración 36 Descarga documentos "form 4" para el año 2006

2007:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2007-10-14 --testing true --overwrite false
```

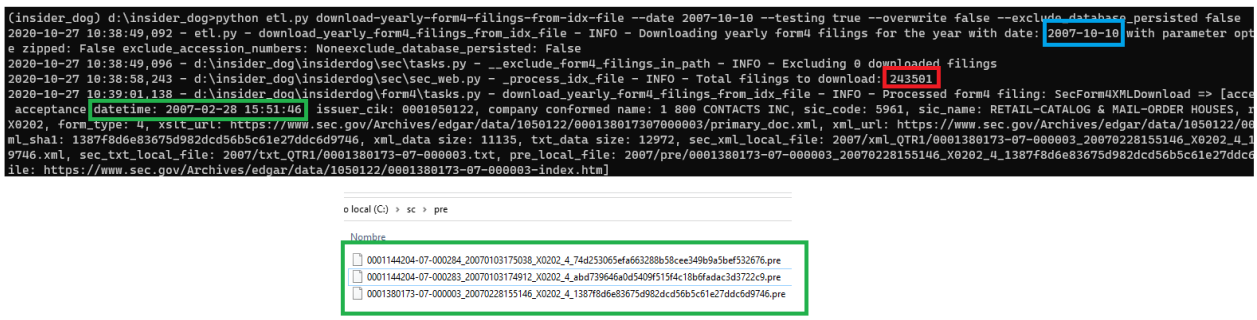


Ilustración 37 Descarga documentos "form 4" para el año 2007

2008:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2008-10-14 --testing true --overwrite false
```

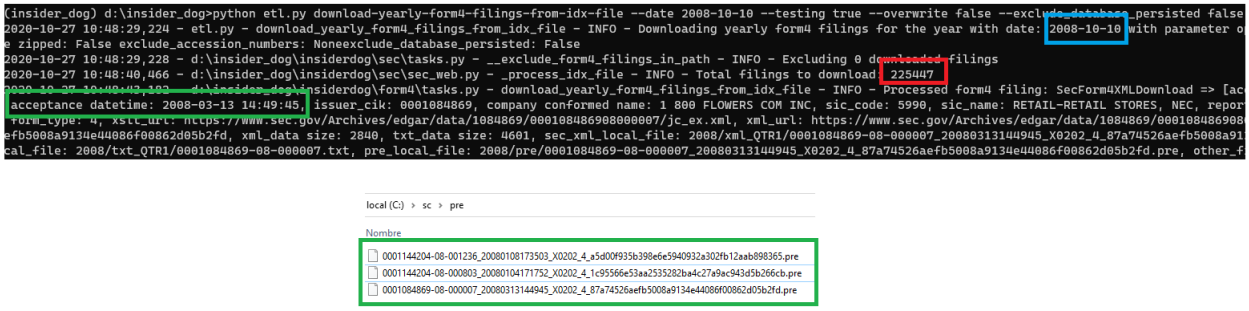
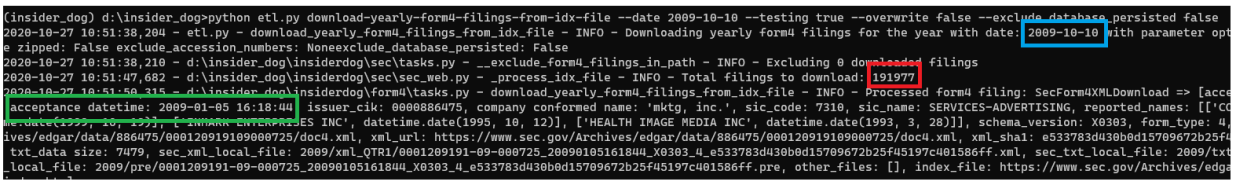


Ilustración 38 Descarga documentos "form 4" para el año 2008

2009:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2009-10-14 --testing true --overwrite false
```



2010:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2010-10-14 --testing true --overwrite false
```

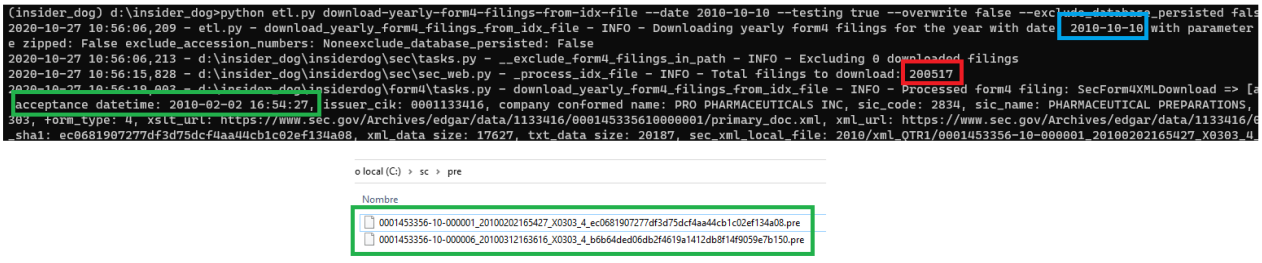


Ilustración 40 Descarga documentos "form 4" para el año 2010

2011:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2011-10-14 --testing true --overwrite false
```

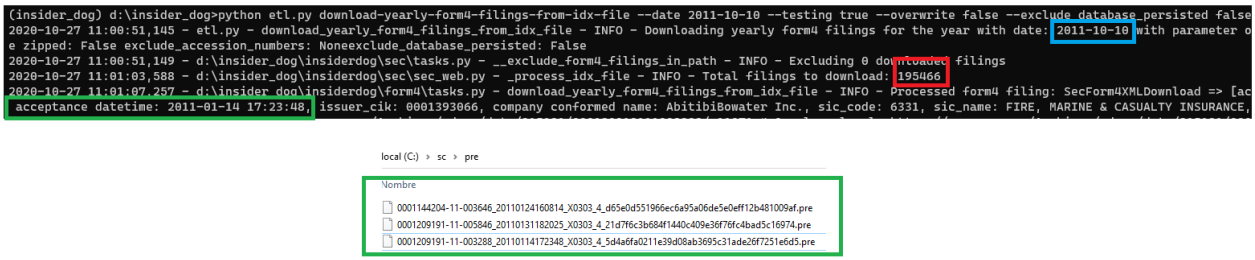


Ilustración 41 Descarga documentos "form 4" para el año 2011

2012:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2012-10-14 --testing true --overwrite false
```

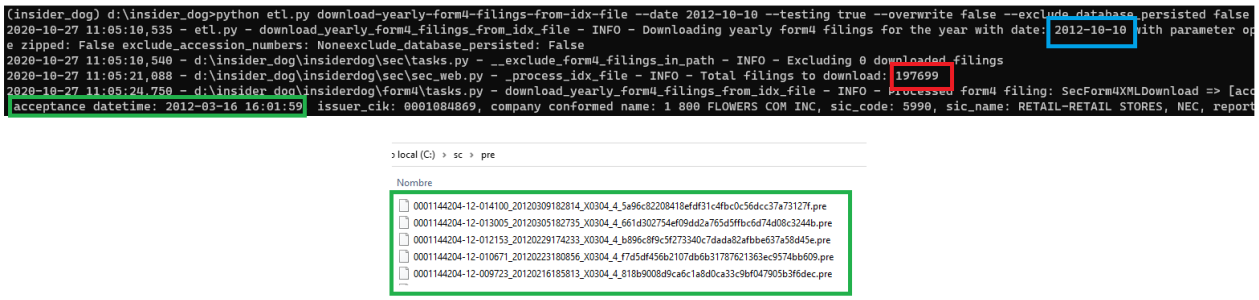


Ilustración 42 Descarga documentos "form 4" para el año 2012

2013:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2013-10-14 --testing true --overwrite false
```

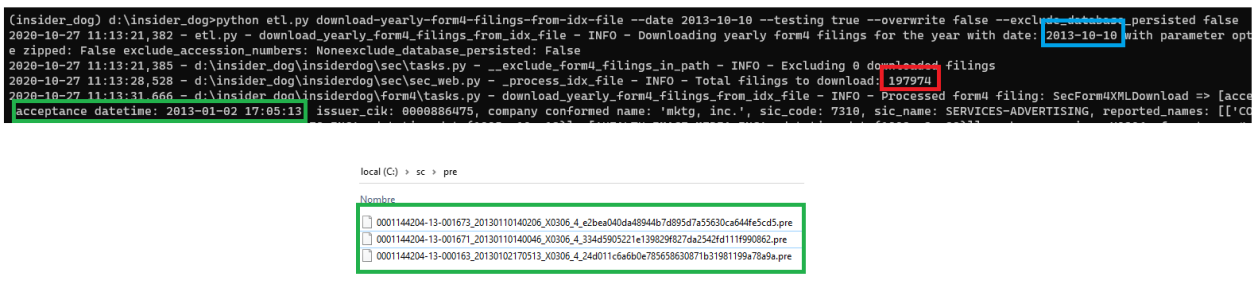


Ilustración 43 Descarga documentos "form 4" para el año 2013

2014:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2014-10-14 --testing true --overwrite false
```

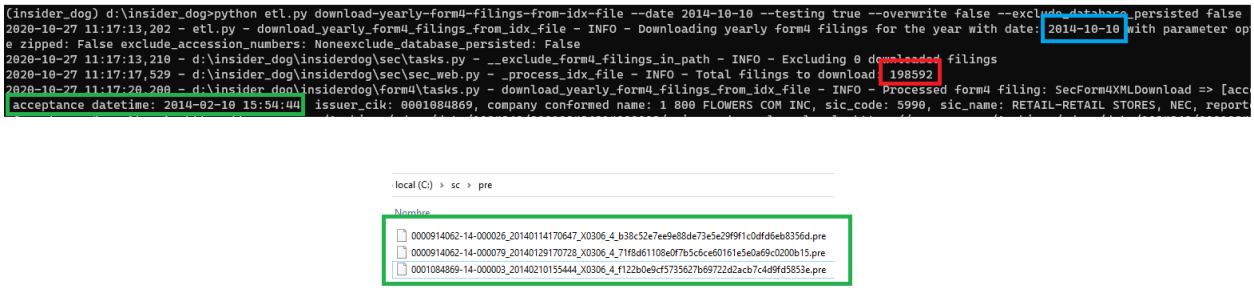


Ilustración 44 Descarga documentos "form 4" para el año 2014

2015:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2015-10-14 --testing true --overwrite false
```

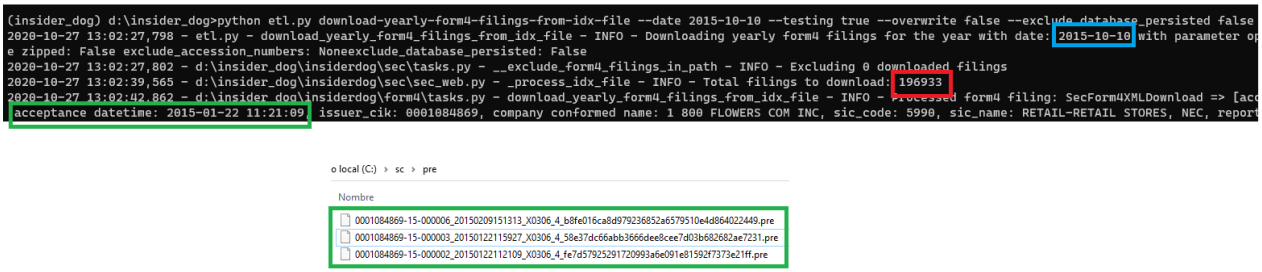
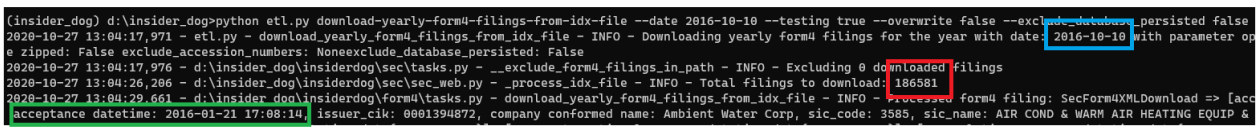


Ilustración 45 Descarga documentos "form 4" para el año 2015

2016:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2016-10-14 --testing true --overwrite false
```



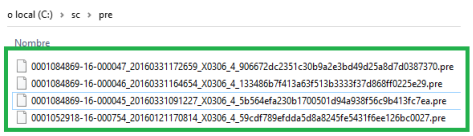


Ilustración 46 Descarga documentos "form 4" para el año 2016

2017:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2017-10-14 --testing true --overwrite false
```

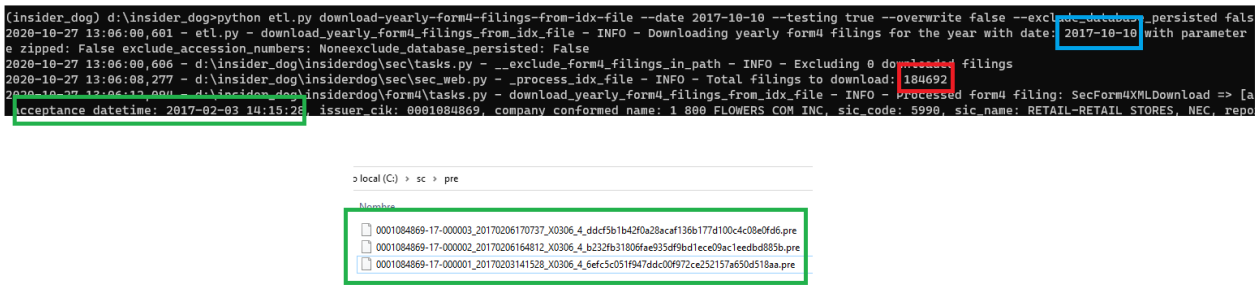


Ilustración 47 Descarga documentos "form 4" para el año 2017

2018:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2018-10-14 --testing true --overwrite false
```

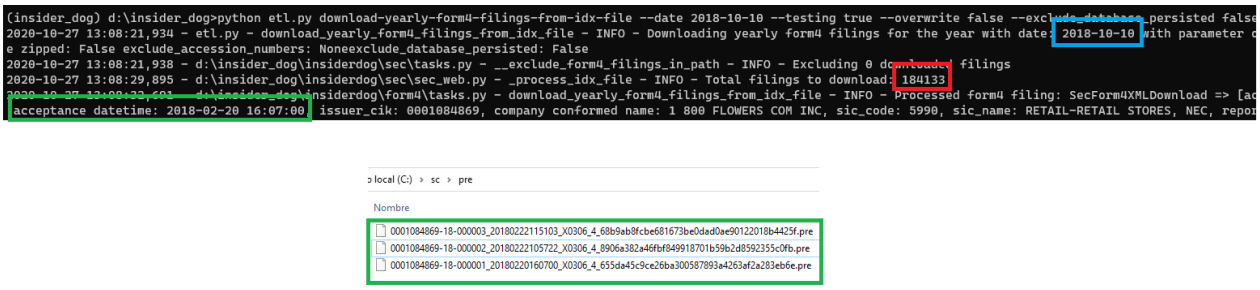


Ilustración 48 Descarga documentos "form 4" para el año 2018

2019:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2019-10-14 --testing true --overwrite false
```

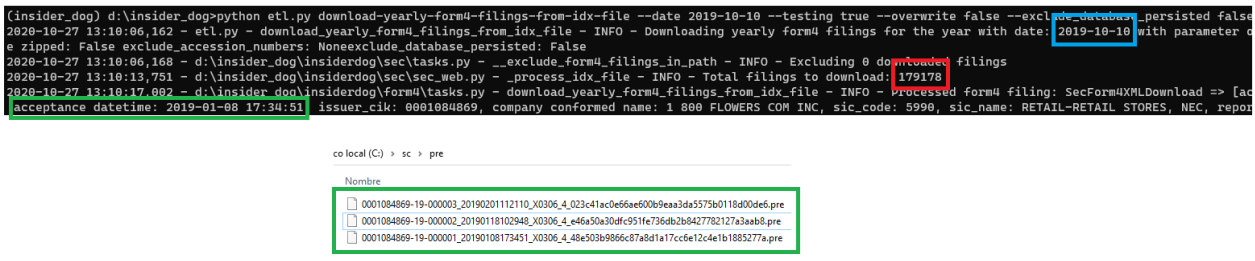


Ilustración 49 Descarga documentos "form 4" para el año 2019

2020:

```
python etl.py download-yearly-form4-filings-from-idx-file --date 2020-10-14 --testing true --overwrite false
```

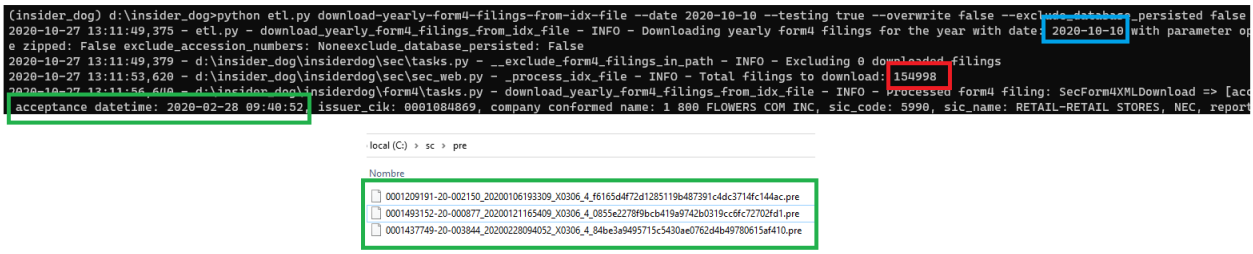


Ilustración 50 Descarga documentos "form 4" para el año 2020

4.2.3 USO DE LA NUBE PARA LA DESCARGA DE DOCUMENTOS *"FORM 4"*

Como se ha expuesto anteriormente, el sistema "EDGAR" debe dar servicio a miles de usuarios simultáneamente y en ocasiones es posible que no esté operativo o la respuesta a una petición de descarga sea lenta.

El número de documentos *"form 4"* a descargar de media en un solo día, es de aproximadamente mil documentos y por lo tanto no se ha previsto en el diseño del software "ETL" el uso operaciones de descarga paralelas, por tratarse de un bajo volumen diario de documentos.

Como se puede ver en la ilustración 51, para descargar el histórico, siempre con la intención de cumplir los plazos establecidos en el proyecto, se han configurado 4 máquinas virtuales "**VPS**" en el proveedor de servicios en la nube Vultr, para que el software "ETL" pueda descargar documentos del sistema "EDGAR" las 24 horas del día, simulando descargas paralelas y controlando las interrupciones de descarga utilizando un script desarrollado en "*bash*".

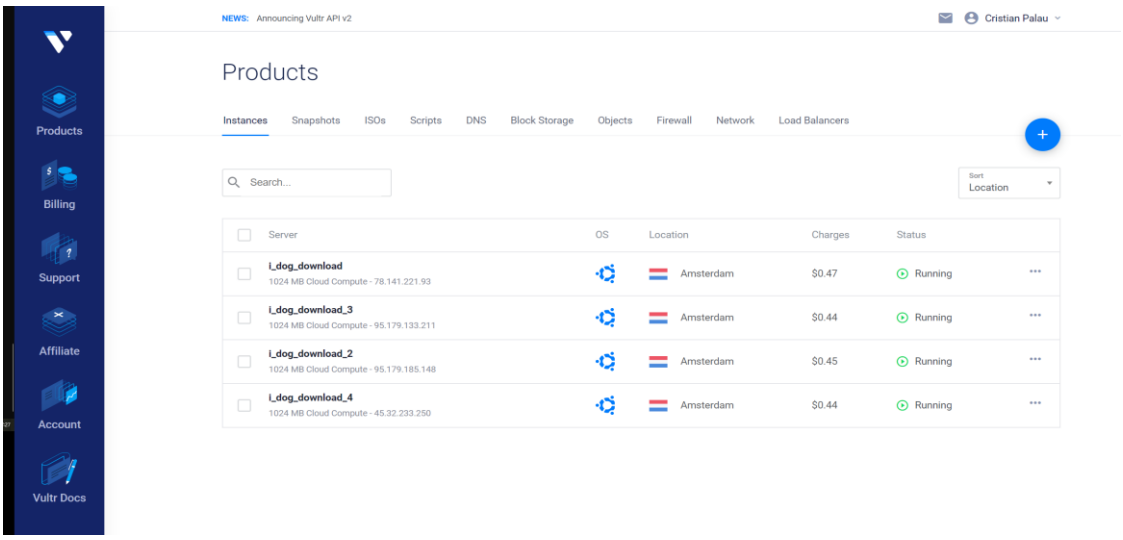


Ilustración 51 Uso del servicio VULTR VPS en la nube para la simulación de descarga paralela de documentos "form 4"

4.2.4 DESCARGA EN TIEMPO REAL UTILIZANDO EL SERVICIO "RSS"

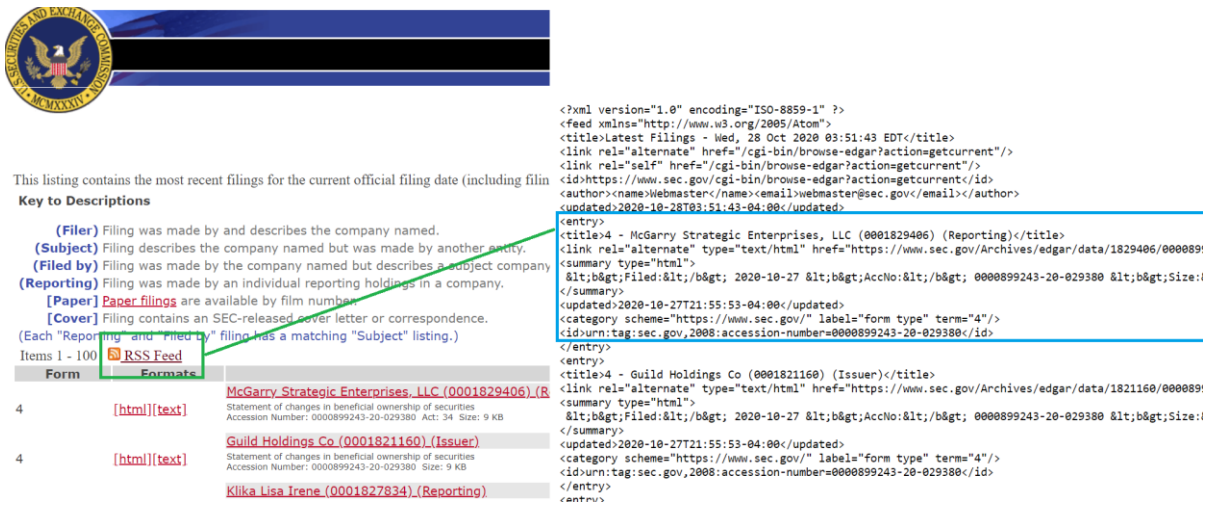


Ilustración 52 Servicio "RSS" de descarga en tiempo real de documentos "form 4" del sistema "EDGAR"

Tal y como se aprecia en la ilustración 52, el sistema "EDGAR" proporciona la opción de subscripción a un servicio "RSS" que permite la descarga en tiempo real de los documentos *"form 4"* que se vayan presentando durante un día.

Este servicio posibilita el análisis de la información en tiempo real y permite al proyecto, en futuras iteraciones, poder analizar los eventos que ocurran de forma inmediata y poder configurar alertas personalizadas.

El software "ETL" desarrollado permite la descarga de documentos *"form 4"* usando el servicio "RSS" de la forma siguiente:

```
python "ETL".py download-sec-rss-form4-filings --testing false
```

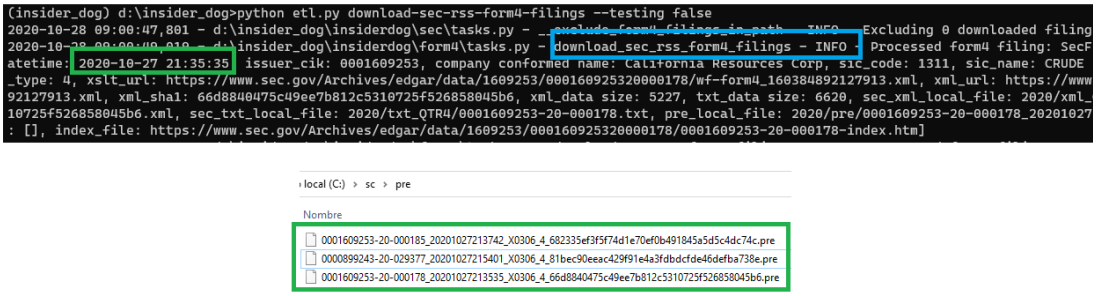


Ilustración 53 Descarga de documentos "form 4" utilizando el servicio RSS

Como se puede observar en la ilustración 53, la salida del comando es idéntica a la de las operaciones de descarga diaria y anual. El software "ETL" desarrollado, hace

uso del principio de reutilización, siendo fácil añadir nuevos tipos de operaciones de descarga en el caso de ser necesario.

4.3 OPERACIONES "ETL" EN EL "*DATA WAREHOUSE* "

4.3.1 Instalación DE LA BASE DE DATOS MARIADB

El proyecto ha sido desarrollado utilizando la base de datos **MariaDB** que es un "*fork*" de **MySQL**, creado por los desarrolladores originales, con una alta compatibilidad binaria y características únicas no presentes en MySQL. La versión utilizada en este proyecto es la versión estable de la rama 10.5, en concreto la versión **10.5.6**. Dado que el proyecto no utiliza características especiales o dependientes de una versión concreta, la instalación podría funcionar perfectamente en las ramas 10.2 o 10.3.

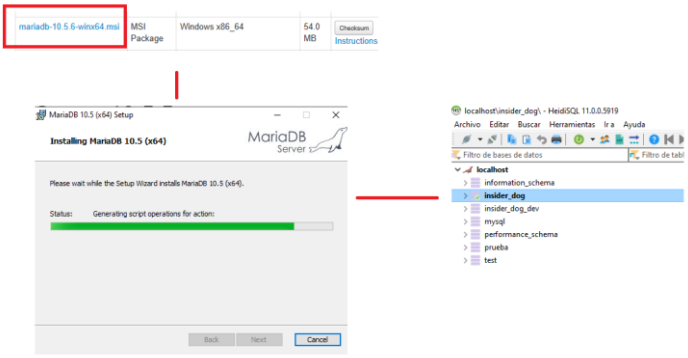


Ilustración 54 Instalación de MariaDB 10.5.6

Para instalar la base de datos, la descargaremos de https://downloads.mariadb.org/mariadb/10.5.6/ y la instalaremos tal como puede verse en la ilustración 54. Es necesario recordar el password del usuario "*root*" que se añada en el momento de la instalación. Para instalaciones de prueba, no es necesario crear un usuario adicional ya que con el usuario "*root*" podemos ejecutar el proyecto y utilizarlo.

Una vez instalada la base de datos, debemos hacer la importación de los datos. Podría utilizarse el software desarrollado para hacer la importación manual de todos los documentos desde el año 2003 hasta la actualidad como puede verse en el punto 4.3.2, pero dado que el proceso de descarga de estos datos es muy lento, se **proporciona para su descarga** una copia de seguridad de la base de datos completa que ha sido utilizada como base para el análisis y por lo tanto esta actualizada al cierre de este proyecto.

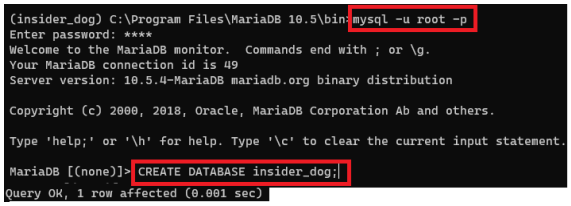


Ilustración 55 Creación de la base de datos "insider_dog"

Para recuperar esta copia de seguridad, primero debemos crear la base de datos "**insider_dog**" accediendo a la consola de MariaDB tal y como se puede ver en la ilustración 55.

Una vez creada la base de datos, procedemos a recuperar la copia de seguridad, descomprimiendo el archivo "**insider_backup.zip**" en un directorio temporal. Una vez descomprimido este archivo, ejecutamos el siguiente comando:

```
mysql -u insider_dog -p insider_dog < insider_backup.sql --show-progress-size
```

El proceso de restauración de los datos puede durar varias horas, dependiendo del hardware sobre el que se ejecute. Aproximadamente puede durar de dos a cinco horas en un equipo con un hardware instalado como el que se ha especificado en este proyecto.

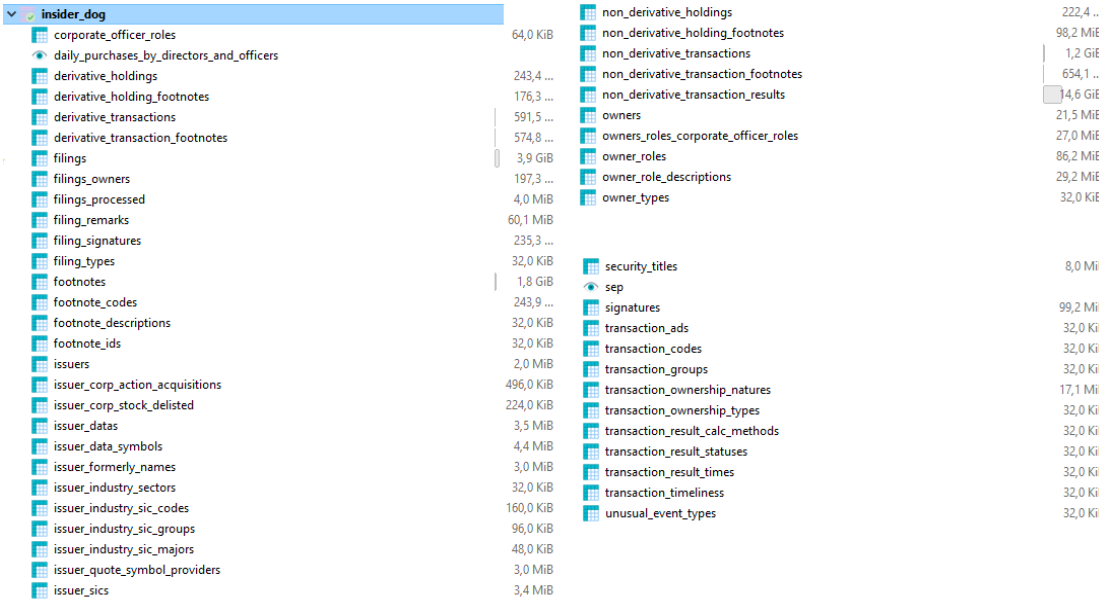


Ilustración 56 Tablas de la base de datos del proyecto

La ilustración 56 muestra las tablas de la base de datos una vez restaurada.

4.3.2 EJECUCIÓN DE OPERACIONES "ETL" SOBRE EL HISTÓRICO DE DATOS

Una vez descargado el histórico de documentos se debe realizar las operaciones "ETL" necesarias para insertar toda la información descargada en el *"data warehouse"*. **Este paso se puede ignorar si se ha recuperado una copia de la BBDD especificado en el apartado anterior.**

A diferencia de la descarga de documentos, que era una característica añadida al software "ETL" desarrollado, la operación "ETL" es la funcionalidad básica del software.

Para la operación "ETL" utilizando el software se debe ejecutar el comando **python etl.py** con los siguientes parámetros base:

- **process-and-etl-xml-pre-filing-file:** Operación que procesa todos los documentos *"form 4"* de un determinado directorio o fichero (parámetro base).

La ejecución del comando acepta los siguientes parámetros:

- **--file_path:** Parámetro que acepta un argumento del tipo string que define un directorio que contenga documentos ".pre" o un determinado archivo a procesar.

- **--testing:** Parámetro que acepta un argumento del tipo (true/false) y que indica si se está ejecutando en un entorno de pruebas.

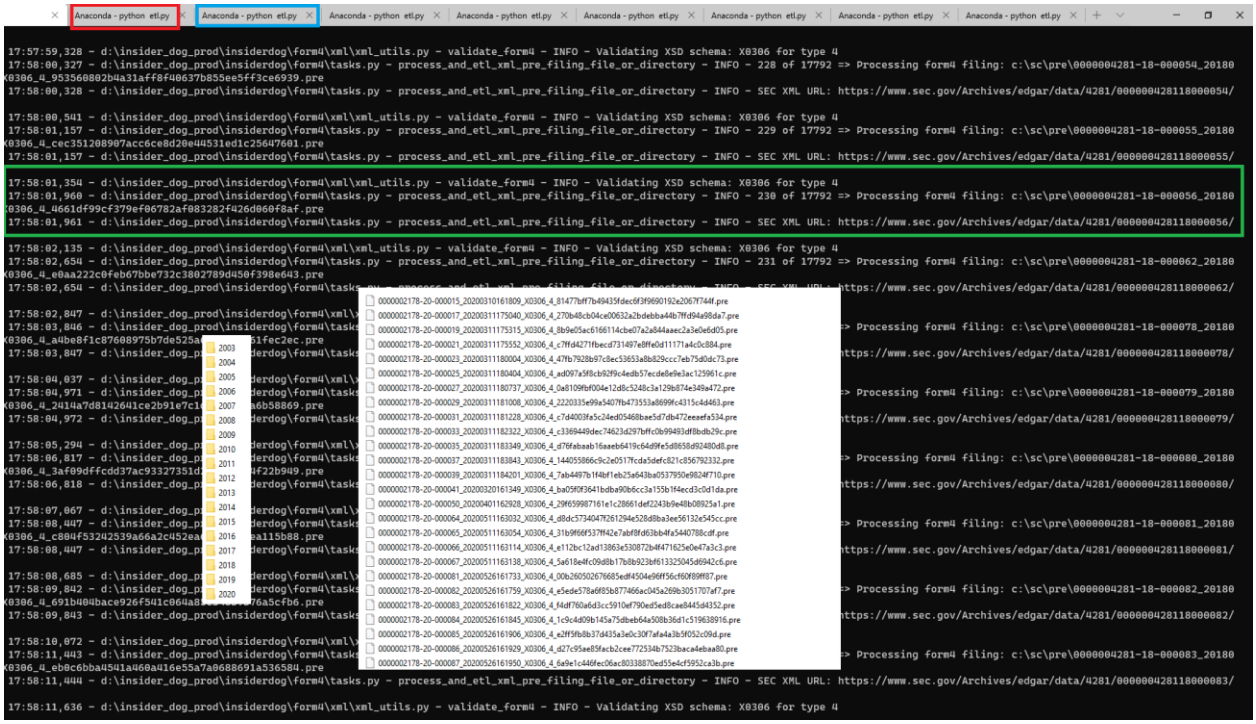


Ilustración 57 Operación "ETL" de proceso de todos los documentos "form 4" (2003-2020) simulando paralelismo

De la misma forma que en el la operación de descarga de documentos, se ha simulado un proceso "ETL" paralelo, ejecutando los siguientes comandos en sesiones de "*shell*" distintas con el objetivo de aprovechar al máximo los recursos del hardware:

2003:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2003\\pre
```

2004:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2004\\pre
```

2005:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2005\\pre
```

2006:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2006\\pre
```

2007:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2007\\pre
```

2008:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2008\\pre
```

2009:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2009\\pre
```

2010:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2010\\pre
```

2011:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path  c:\\data\\2011\\pre
```

2012:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2012\\pre
```

2013:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2013\\pre
```

2014:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2014\\pre
```

2015:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2015\\pre
```

2016:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2016\\pre
```

2017:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2017\\pre
```

2018:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2018\\pre
```

2019:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2019\\pre
```

2020:

```
python etl.py process-and-etl-xml-pre-filing-file --file_path c:\\data\\2020\\pre
```

Una vez ejecutadas y finalizadas las operaciones "ETL" sobre el histórico de datos, únicamente falta un paso para poder usar el *"data warehouse"* como herramienta analítica, la limpieza de datos.

El software "ETL" como entregable se puede descargar en este enlace.

4.4 LIMPIEZA Y NORMALIZACIÓN DE DATOS "*DATA CLEANING*"

Cuando se procesa información de terceros, la probabilidad de que estos datos estén normalizados es mínima, siendo necesario la aplicación de operaciones de limpieza de datos *"**data cleaning**"* para que puedan aportar valor analítico.

Como veremos en el apartado de análisis de datos, un dato muy importante para las conclusiones del proyecto es el rol del *insider* dentro de la compañía. A nivel analítico resulta muy interesante conocer si una transacción de compra la ha realizado el *"**Chief Financial Officer**"* (director financiero) o un vice presidente ejecutivo.

Desde un punto de vista de valor aportado en la toma de una decisión financiera es muy importante ya que ofrece un factor de credibilidad en la operación reportada.

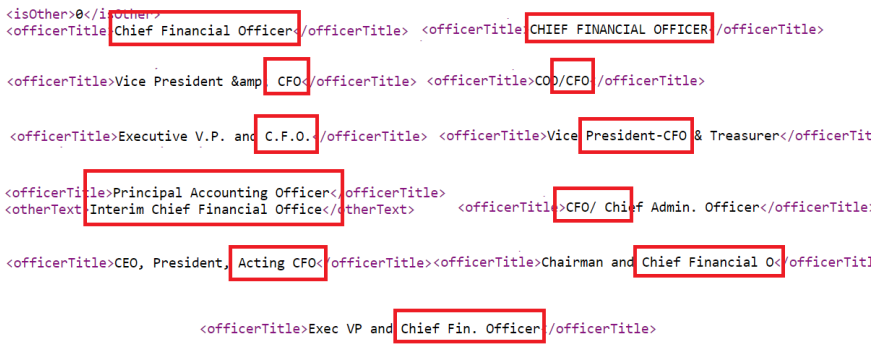


Ilustración 58 Diferentes formas de notificar el rol de director financiero (CFO) en la que se aprecia la falta de normalización del dato.

La ilustración 58 muestra diferentes formas de notificar el rol de director financiero (**CFO**) en los datos "XML" y muestra claramente el problema anteriormente expuesto. El sistema "EDGAR" a través del esquema "XML" no impone a la empresa la forma de notificar este dato a través de una lista predefinida de roles, por lo que es necesario realizar una operación de limpieza y normalización de la información almacenada en el *"data warehouse"* sobre este dato.

4.4.1 USO DE OPEN REFINE PARA LA LIMPIEZA DE DATOS

Para la limpieza y normalización de los roles se ha utilizado la herramienta Open Refine que fue liberada por Google en licencia de código abierto. Esta herramienta, permite la limpieza de grandes volúmenes de datos aplicando diferentes técnicas que permiten su normalización.

Una vez descargada de su página Web oficial, se procederá a descomprimir en un directorio de la unidad C, llamado "**c:\openrefine-3.3**".

Para poder limpiar los datos usando la herramienta, debemos primero exportar en formato "CSV" toda la lista de roles con su clave primaria que el proceso "ETL" ha importado previamente, para posteriormente una vez normalizados, realizar la importación de los datos en el *"data warehouse"*. Como se puede ver en la ilustración 56 la "***query***" SQL genera un archivo "CSV" que permitirá la importación en la herramienta.

```
SELECT ore.id, orde.`data` FROM owner_roles ore, owner_role_descriptions orde
WHERE orde.id = ore.officer_title_id
ORDER BY ore.id ASC
INTO OUTFILE 'c:\\sc\\roles.csv'
FIELDS ENCLOSED BY '"'
TERMINATED BY ';'
ESCAPED BY '"'
LINES TERMINATED BY '\r\n';
```

id	data
1	Executive Vice President
2	Chief Marketing Officer
3	Vice President, Emea
4	Treasurer
5	Chief Accounting Officer
6	President And Ceo
8	President, Global Development
9	Vice President And Coo
10	President And Ceo
11	President - The Marmaxx Group
12	Sr. Vice President, Technology
13	Vp Of Special Projects
14	Cfo & Executive Vp

Ilustración 59 Exportación de la información de roles para su limpieza y normalización en Open Refine

Una vez se dispone del fichero “CSV”, ejecutaremos Open Refine con el comando: “**C:\openrefine-3.3\openrefine.exe**” y crearemos un proyecto en Open Refine y lo importaremos tal y como muestra la ilustración 60.

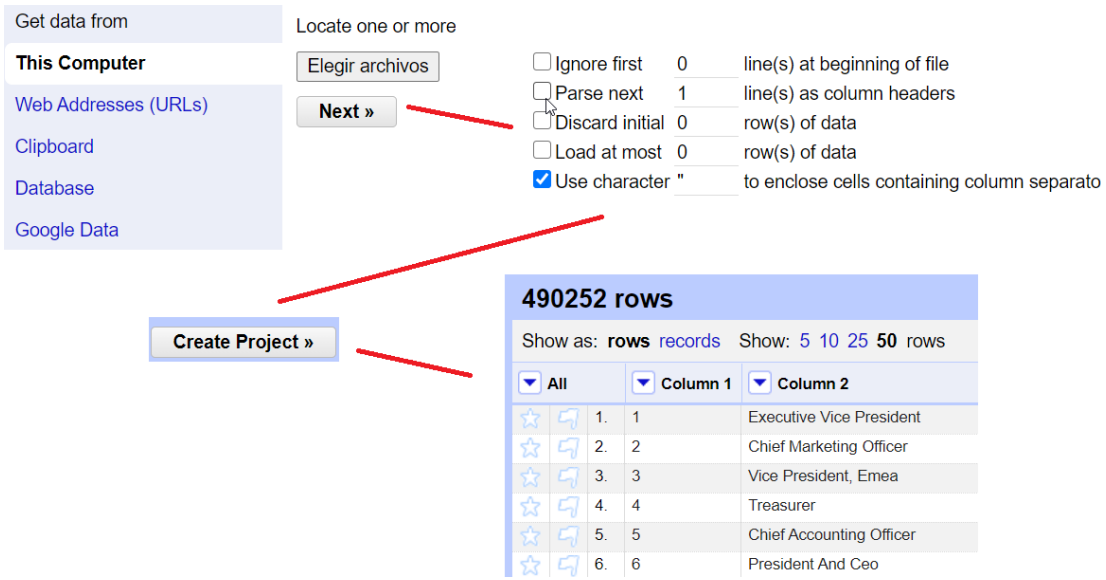


Ilustración 60 Importar un archivo “CSV” en Open Refine y crear un proyecto

Open Refine es una herramienta compleja y su uso queda fuera del ámbito de este proyecto, sin embargo, en la ilustración 61 podemos ver un ejemplo de una de sus funciones que es la limpieza agrupada del rol “**President And Ceo**” por la versión normalizada “**President & Chief Executive Officer**”. Esta sustitución permite poder normalizar los roles de “**President**” y “**Chief Executive Officer**” aplicando esta modificación a 6777 filas de todo archivo CSV exportado.

En el proceso de normalización de estos datos se conserva la clave primaria original de la tabla “**owner_roles**” que permitirá relacionar posteriormente el nuevo rol en el proceso de importación en el *“data warehouse”*.

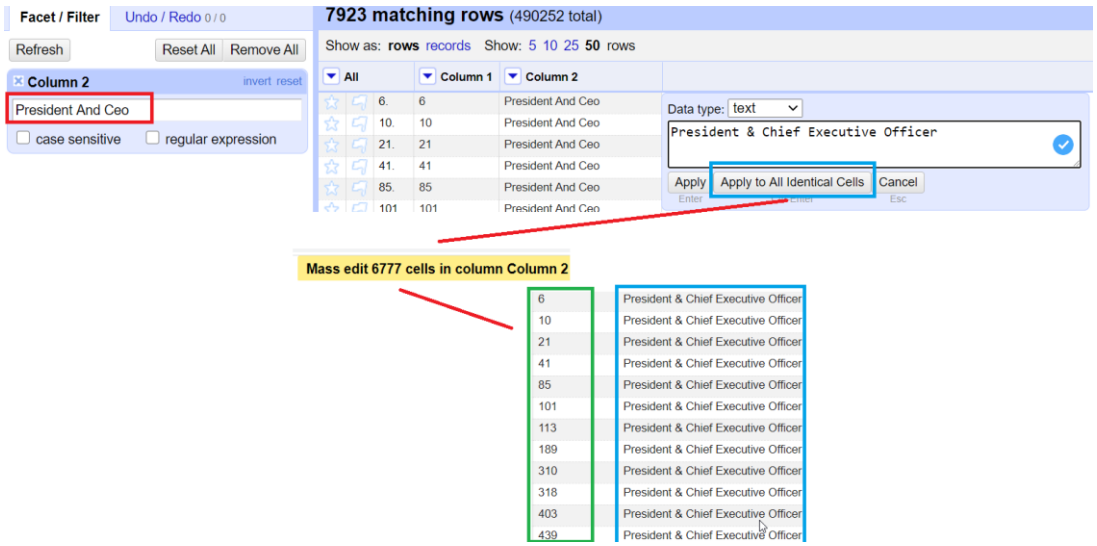


Ilustración 61 Normalización de un rol utilizando una de las funciones de Open Refine

Este proceso de normalización tiene un coste muy alto en horas ya que se deben normalizar **490.252** roles.

Una vez aplicada toda la normalización, se exportará de nuevo un fichero "CSV" desde Open Refine, con los roles normalizados.

Para la importación del trabajo de normalización, se utilizará un script desarrollado en el lenguaje Ruby llamado "**import_roles.rb**" disponible en la carpeta "**scripts**" del software "ETL". Este script realizará la importación en el "data warehouse". Una vez importado se muestra a través de una "***query***" SQL el resultado de la normalización de los datos, mostrando que para una misma clave primaria ID en la tabla "**owner_roles**" disponemos de los dos roles normalizados. En la ilustración 62 se puede observar todo el proceso de importación.

Como no se realizarán de forma periódica normalizaciones de rol, se ha decidido no incluir esta funcionalidad como opción dentro del software "ETL" optando por el desarrollo de un script.

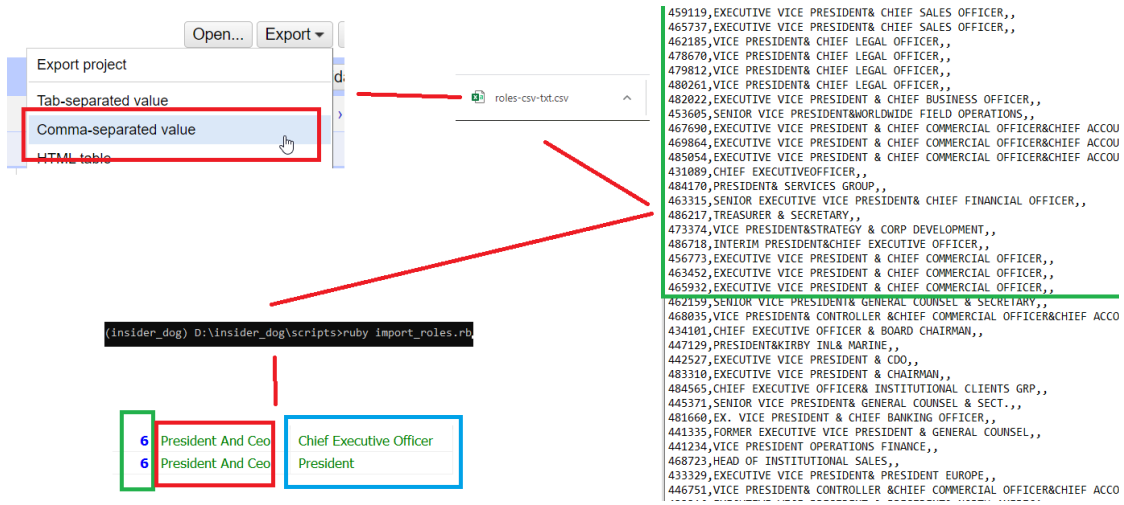


Ilustración 62 Importación de roles normalizados en el "data warehouse"

4.5 *"DATA WAREHOUSE"*

Una vez concluidas las operaciones "ETL" sobre todo el histórico de datos y la normalización de los roles de los *"insiders"*, el siguiente paso es el uso del "*data warehouse"* resultante como herramienta para el análisis de datos.

Antes de proceder al análisis de datos, es necesario comprobar que si solicitamos la información de un documento *"form 4"*, el sistema nos proporcionará con exactitud toda la información de ese documento, concluyendo así que el *"data warehouse"* cumple con los requerimientos especificados en la fase de diseño.

La ilustración 63 muestra un documento con "***accession number***" 0001037868-03-000004 y su correspondiente representación de esos datos extraídos del *"data warehouse"*.

4.5.1 REPRESENTACIÓN DE UN DOCUMENTO *"FORM 4"* EN EL *"DATA WAREHOUSE"*

*Ilustración 63 Representación en el "data warehouse" del documento "form 4 XML" con "accession number" **0001037868-03-000004***

4.5.2 INFORME FINAL DEL *"DATA WAREHOUSE"*

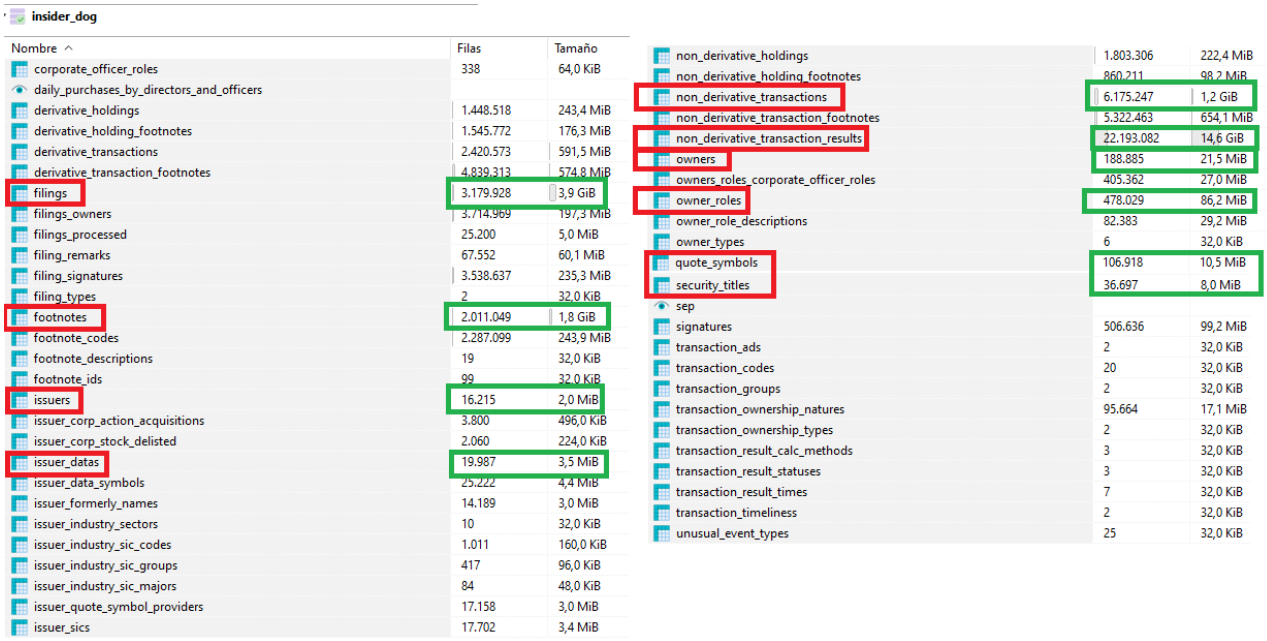


Ilustración 64 Tablas, registros y tamaño físico del "data warehouse"

En la ilustración 64 podemos ver las tablas, registros y tamaño físico del *"data warehouse".* El tamaño aproximado total es de 25 Gb.

Tal y como se muestra en la ilustración, a cierre del proyecto se destacan las siguientes entidades importadas:

- Documentos *"form 4"* ("***filings***"): 3,1 millones de registros.
- Empresas ("***Issuers***"): 16.215 registros con 19.987 descripciones ampliadas y 106.918 símbolos de cotización.
- Directivos (***"insiders"***): 188.885 registros con 478.029 roles de empresa registrados.
- Operaciones reportadas ("***Non Derivativative Transactions***"): 6,1 millones de operaciones de las cuales **824.817** son operaciones de compra de acciones, objetivo de análisis del proyecto.
- Notas de transacción ("***Footnotes***"): 2,01 millones de notas descriptivas de operación reportada.

Se ha definido una vista en la base de datos, denominada "**Daily_purchases_by_directors_and_officers**" que proporciona el total diario de operaciones de compra reportadas por los *"insiders"*.

5 ANÁLISIS DE DATOS Y CONCLUSIONES

5.1 USO DEL ENTORNO JUPYTER PARA EL ANÁLISIS DE DATOS

Para el análisis de datos, el proyecto utiliza **Pandas**, librería desarrollada en Python que facilita la manipulación y el análisis de datos, así como varios scripts desarrollados para la extracción de datos. Pandas se ha utilizado principalmente para el análisis exploratorio de los datos y aunque ya se ha instalado previamente, en esta sección se explica brevemente como comprobar que funciona correctamente junto a Jupyter, que es un entorno también muy utilizado por los científicos de datos para analizar "***datasets***". Jupyter se basa en el intérprete de Python "**IPython**" y permite, entre otras cosas, la ejecución selectiva de líneas de código, facilitando que el cambio en una línea sea inmediatamente resuelto por el intérprete.

Al instalar el software Anaconda, disponemos del entorno Jupyter totalmente configurado y listo para funcionar.

Como podemos ver en la ilustración 65, para poder ejecutar Jupyter debemos ejecutar nuevamente la aplicación de Anaconda, tal y como hemos hecho en la sección de instalación de librerías Python y seleccionar la opción "***Launch***" en la sección del aplicativo que indica "***Jupyter Notebook***".

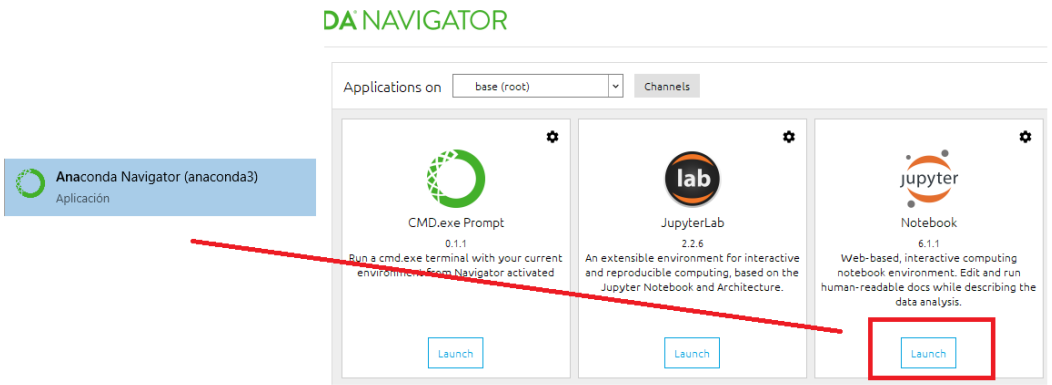


Ilustración 65 Ejecución de Jupyter Notebook

Una vez ejecutado, Jupyter abrirá una pestaña en nuestro navegador en la que podremos establecer nuevas sesiones en el intérprete.

Para establecer una nueva sesión debemos hacer clic en la opción "***New***" del botón que este situado en la parte superior derecha y seleccionar "*Python 3*" para iniciar una nueva sesión en el intérprete como se muestra en la ilustración 66.

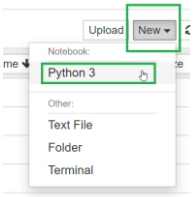


Ilustración 66 Ejecución de una nueva sesión en Jupyter

Una vez abierta una sesión Python en Jupyter, ya podemos ejecutar código en Python.

Para comprobar que tenemos Pandas correctamente instalado ejecutaremos la instrucción que se muestra en la ilustración 67. Si podemos observar el número de versión, tenemos la instalación del entorno totalmente configurada.

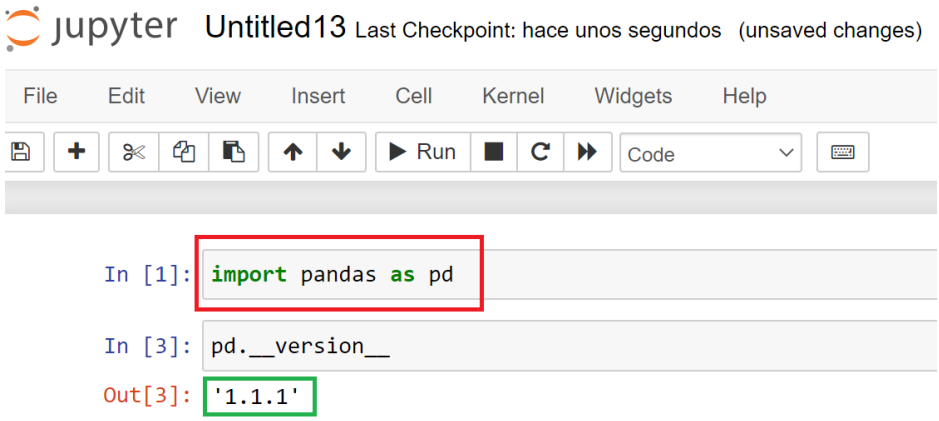


Ilustración 67 Sesión en Jupyter Notebook usando Pandas

5.2 ANÁLISIS EXPLORATORIO INICIAL

La librería **Pandas** permite importar datos para su análisis en diferentes formatos como "CSV" o "Excel" así como importar datos directamente desde una "***Query***" o tabla en la base de datos.

Por su diseño interno, la librería realiza todas las operaciones en memoria y cuando las fuentes de datos superan los 2 Gb de tamaño, su uso es lento y en ocasiones las operaciones no pueden finalizarse por falta de memoria. Como el *"data warehouse"* desarrollado supera este tamaño máximo, se ha optado por proporcionar al proyecto tres ficheros en formato "**CSV"** extraídos del *"data warehouse"* que se pueden obtener en este enlace y que contienen un listado de las operaciones reportadas de forma desnormalizada, así como la lista de compañías y *"insiders"*. El total del "***dataset***" es de 1,2 Gb de tamaño, un tamaño con la que Pandas puede trabajar sin grandes problemas de rendimiento o consumo excesivo de memoria.

Se ha optado por el uso de los ficheros "CSV" para poder indicarle a Pandas el tipo de dato y el tamaño en bytes de cada campo, con el objetivo de ahorrar el máximo de memoria posible. Pandas, normalmente suele inferir el tipo de dato, pero en ficheros de tipo texto, como los "CSV", es recomendable indicar el tipo y tamaño del dato para obtener el máximo rendimiento durante el análisis.

Para el análisis exploratorio inicial también se hará uso de la librería **Matplotlib** que permite representar los datos con gráficos y que está incluida por defecto al realizar la instalación de la librería Pandas.

Para poder empezar con el análisis, dentro de una sesión en Jupyter ejecutaremos la importación de datos a Pandas como se puede ver en la ilustración 68:

```
import pandas as pd
```

```
issuers = pd.read_csv('D:\\issuers.csv', sep=';',lineterminator='\n', header=0,
                      names=['issuer_cik', 'name', 'setup_date'], parse_dates=["setup_date"],
                      dtype={'issuer_cik': 'category', 'name': 'str'})
```

```
# Ordenamos por fecha de creación y eliminamos los duplicados
issuers = issuers.sort_values('setup_date', ascending=False).drop_duplicates(subset=['issuer_cik'], keep='first')
```

```
owners = pd.read_csv('D:\\owners.csv', sep=';',lineterminator='\n', index_col=False, header=1,
                      names=['owner_cik', 'name', 'code'],
                      dtype={'owner_cik': 'str', 'name': 'str', 'code': 'category'})
```

```
# especificamos los tipos y añadimos que interprete \\N como NaN
df = pd.read_csv('D:\\t.csv', sep=';', lineterminator='\n', index_col=False,
                 names=['date', 'accession_number', 'accession_time','filing_type','issuer_cik',
                        'shares','price', 'shares_owned_post_trans', 'tc_code', 'ads', 'ownership_type',
                       'owner_cik','is_director','is_officer','is_ten_percent', 'is_other',
                        'roles','is_10b5_1','is_16_b','is_16_a11'],
                 parse_dates=["date", "accession_time"],
                 na_values=['\\N'],
                 dtype={'tc_code': 'category', 'ads': 'category', 'ownership_type': 'category', 'filing_type': 'category',
                       'is_director': 'bool', 'is_officer': 'bool', 'is_ten_percent': 'bool', 'is_other': 'bool',
                       'issuer_cik': 'category', 'owner_cik': 'str', 'roles': 'category', 'accession_number': 'str',
                       'shares':'float32', 'price': 'float32', 'shares_owned_post_trans': 'float32',
                        'owner_cik': 'category', 'is_10b5_1': 'bool', 'is_16_b':'bool', 'is_16_a11':'bool'})

total_days_passed = (df['accession_time'] - df['date']).dt.days
df.insert(3, column='days_past', value=total_days_passed)

total_value = df['shares'] * df['price']
df.insert(7, column='total_value', value=total_value)

is_stock_option = (df['is_10b5_1'] | df['is_16_b'] | df['is_16_a11'])
df.insert(18, column='is_stock_option', value=is_stock_option)

df = df.drop(['is_10b5_1','is_16_b','is_16_a11'], axis=1)
```

```
pt = df[df['tc_code'] == 'P']
```

Ilustración 68 Importación de datos "CSV" en Pandas para el análisis exploratorio

Toda la sesión del análisis exploratorio en Jupyter se puede descargar de este enlace. Esta importación inicial es la base para poder realizar las exploraciones iniciales.

Es importante recordar, que el análisis exploratorio trabaja con las operaciones reportadas, que son aproximadamente unos 6,1 millones y no a nivel de documento *"form 4"*. Cada documento *"form 4"* puede tener N transacciones, y cada transacción puede pertenecer a varios *"insiders"* por lo que, como veremos en el análisis exploratorio, una vez desnormalizados los datos, el total de registros de operación asciende a aproximadamente **7 millones de operaciones**.

Por razones de espacio, en el análisis exploratorio, en la mayoría de las secciones no se incluye el código Python que ha generado el gráfico de la ilustración.

Los datos del año 2020, no están completos al cierre del proyecto.

5.2.1 ERRORES EN LA INFORMACIÓN SUMINISTRADA AL SISTEMA "EDGAR"

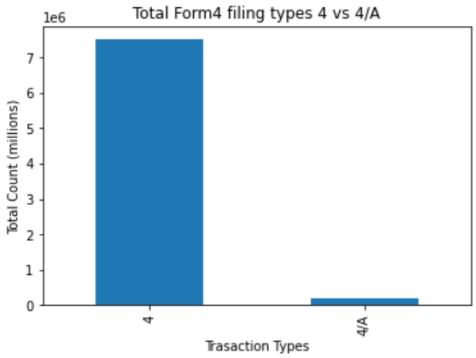
Uno de los aspectos fundamentales en cualquier análisis de datos es determinar qué grado de fiabilidad tienen los datos con los que se trabaja. A pesar de que el sistema "EDGAR", como se ha visto anteriormente, obliga a las empresas a comunicar los datos en un determinado formato y estructura, sabemos que por la propia definición del

sistema, existe un tipo de documento, denominado ***"form 4/A"***, que las empresas utilizan para corregir un documento anterior enviado con errores.

También es importante saber cuántas operaciones de compra han sido presentadas y que todavía no se han producido, un error que necesitamos conocer si se produce con frecuencia, ya que indica un fallo al reportar la fecha de una transacción de compra y por lo tanto la no validez de la operación.

```
df.groupby(['filing_type']).size().sort_values(ascending=False).plot(kind="bar", xlabel='Trasaction Types',
                                                                     ylabel='Total Count (millions)',
                                                                     title='Total Form4 filing types 4 vs 4/A')
```

```
<AxesSubplot:title={'center':'Total Form4 filing types 4 vs 4/A'}, xlabel='Trasaction Types', ylabel='Total
```




```
# detectar erroes como fechas de presentación anteriores a la fecha de la transacción
(pt['days_past'] < 0).sum()
```

```
56
```

Ilustración 69 Análisis de errores en la información suministrada al sistema "EDGAR"

Como podemos ver en la ilustración 69, el nivel de transacciones reportadas como error es muy bajo (documentos del tipo 4 vs tipo 4/A), así como el número de operaciones de compra que hay que descartar por no haberse producido, que son sólo **56**. Aunque es imposible saber si la información reportada es 100% fiable, ya que la única conocedora de su fiabilidad es la empresa, sabemos que el porcentaje de error al reportar esta información es muy bajo y por lo tanto **podemos concluir que las empresas suelen realizar un buen trabajo al comunicar la información**.

5.2.2 HORA Y DÍA DE LA SEMANA

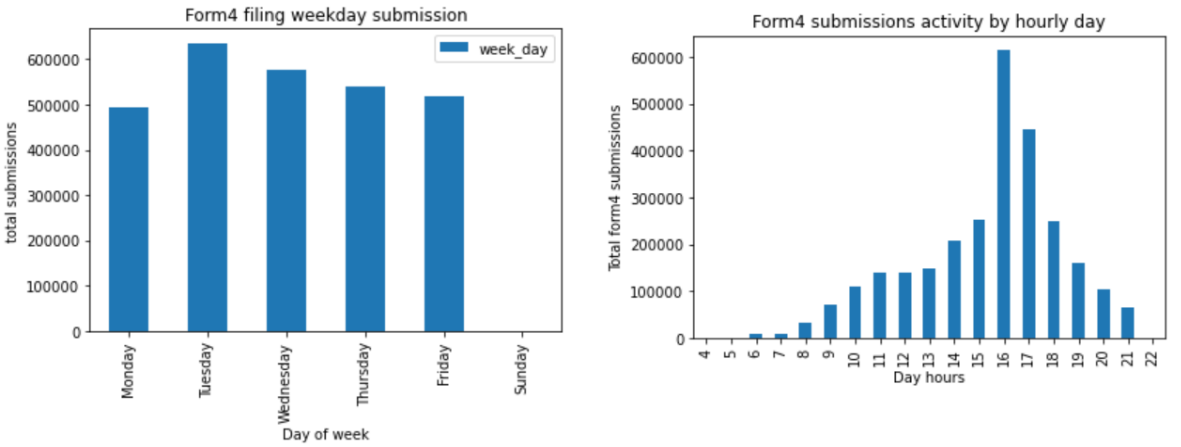


Ilustración 70 Análisis del día de la semana y la hora más frecuente en la presentación de los documentos "form 4"

Un dato importante para un inversor que quiera realizar un seguimiento en tiempo real de las operaciones de compra reportadas, es el día de la semana y sobre todo a qué hora se publican la gran mayoría de operaciones. Como podemos ver en la ilustración 70, **el martes es el día en que hay un mayor número de envíos al sistema "EDGAR"** (aparece el domingo con algún envío pero se debe a errores del propio sistema) y también podemos ver el primer dato relevante para el análisis que es el referente a la hora de la presentación. Como podemos ver, **la mayoría de las transacciones se han reportado a las 16:00 horas, una vez ha cerrado el mercado bursátil**. Este dato, penaliza al inversor que quiera mimetizar la compra de una determinada acción, ya que al comunicarse después del cierre del mercado, será muy complicado obtener el mismo precio que pago el "*insider*" por ella.

5.2.3 DOCUMENTOS "FORM 4" Y TIPOS DE OPERACIÓN

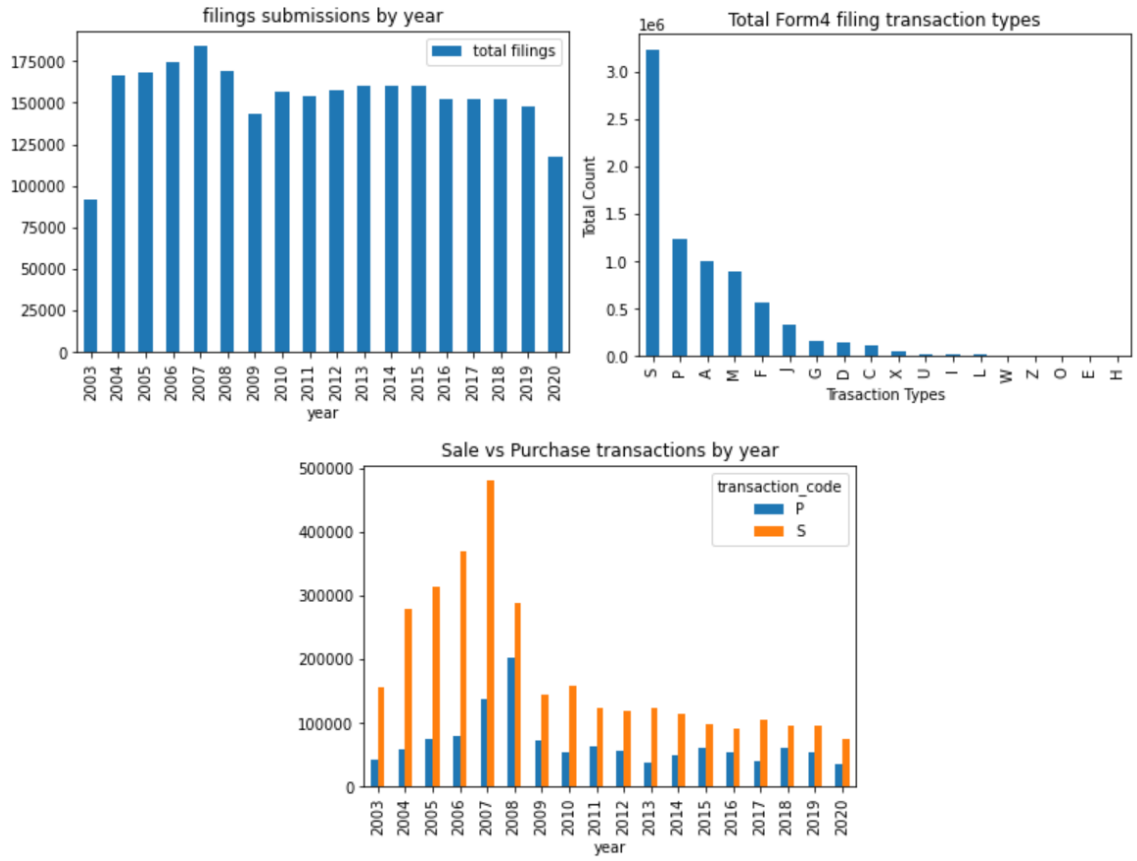


Ilustración 71 Análisis de los documentos "form 4", operaciones presentadas por año y tipo de operaciones más reportadas

La ilustración 71 muestra que el tipo de operación más reportada, es la de venta de acciones (S). Como se puede observar, el volumen de las operaciones de venta es tres veces superior a las operaciones de compra. Esto hecho, proporciona una conclusión importante: **Los *"insiders"* son fundamentalmente vendedores de acciones**. Por este motivo, cuando realizan una compra, objetivo principal de este proyecto, hay que analizarla en detalle porque es un hecho significativo.

En la ilustración también se observa que el número de documentos *"form 4"* presentado cada año ha ido disminuyendo desde el inicio de la publicación de la información en formato "XML" o también podría interpretarse que en momentos de grandes crisis económicas, los *"insiders"* aceleran sus operaciones, presentando más documentación sobre sus operaciones.

Por último, podemos ver por cada año del histórico de datos, el porcentaje de operaciones de venta versus compra. Es muy significativo, que **en el año previo a la finalización de un mercado bajista bursátil se aceleré la compra de acciones** tal y como vemos en la reducción de la divergencia de compra-venta en el año 2008, que fue previa al suelo de mercado del año 2009.

5.2.4 *"INSIDERS", "OWNERS"* Y ROLES

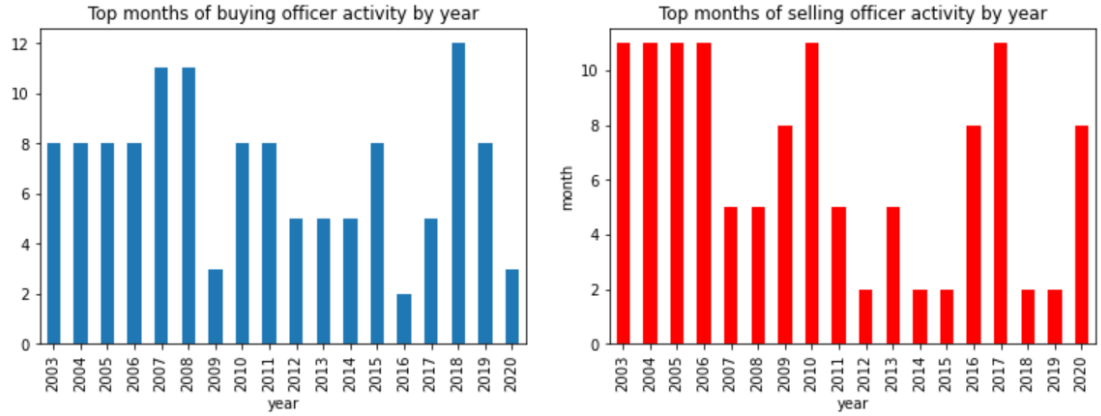


Ilustración 72 Análisis de los meses del año en el que se más se producen operaciones de compra o venta

La ilustración 72 proporciona un dato esencial para cualquier inversor que desee realizar un seguimiento de las operaciones de compra con el objetivo de establecer posiciones en el mercado, ya que indica que **la mayoría de operaciones de compra se realizan en el entre el tercer y cuarto trimestre del año (agosto-diciembre)**. Los gráficos muestran solamente aquellas operaciones realizadas por los ejecutivos que tienen cierto peso en el día a día de las operaciones dentro de la empresa (CEO, CFO, etc.) y que son fundamentalmente las operaciones que se analizarán en el proyecto

En la ilustración 73 podemos observar que **la persona física es el tipo de *"insider"* que más operaciones realiza** y es una buena noticia para el análisis ya que son mucho más fiables las decisiones tomadas por una persona que tiene un peso específico en la compañía que empresas con participaciones en el accionariado o fundaciones.

Este dato viene también confirmado en la ilustración que muestra que se deben ignorar aquellas operaciones que no son realizadas por *"insiders"* y también vemos que el tipo de rol que **reporta más operaciones es el director pero no es un rol que proporcione gran fiabilidad como el cargo ejecutivo**. (Moreland, J. (2000). Profit from Legal Insider Trading. p. 52)

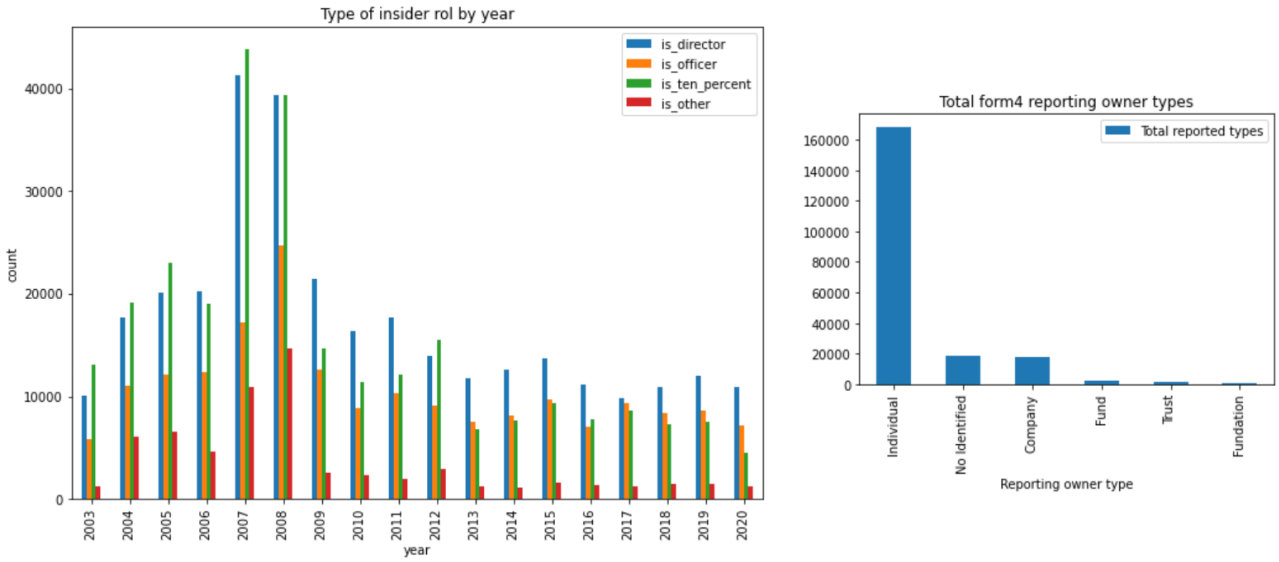


Ilustración 73 Análisis de los roles de los "insiders" más frecuentes por año y tipo de "insider" más frecuente.

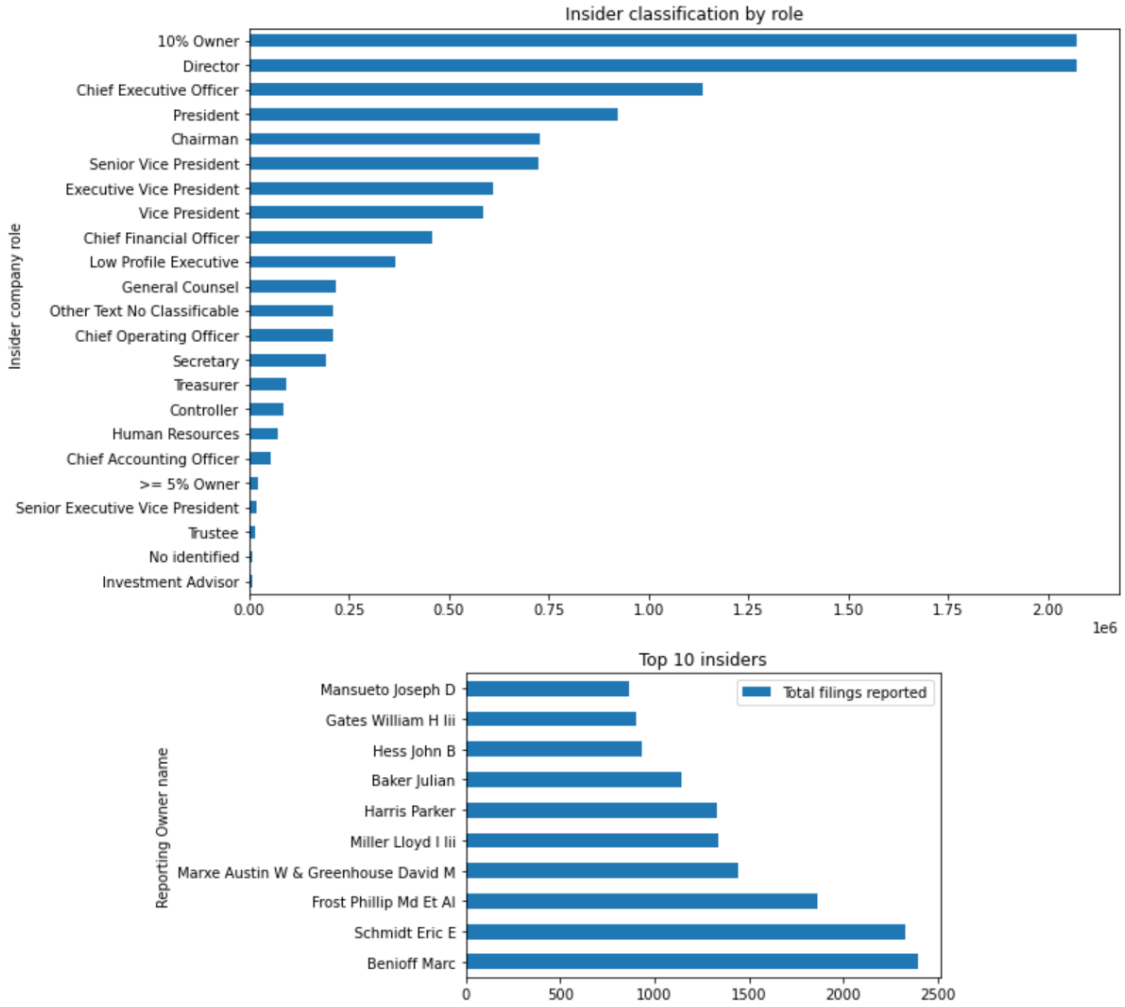


Ilustración 74 Análisis del rol más frecuente y top 10 de los "insiders" más activos

También se muestra en la ilustración 74 el top 10 de los *"insiders"* que más operaciones realizan y como curiosidad podemos ver nombres tan ilustres como Gates William H III, conocido como **Bill Gates,** ex CEO de Microsoft o **Eric S. Schmidt** ex CEO de Google, Inc**.** A pesar de que estos nombres, podrían llevar a pensar que es gente con gran conocimiento del mercado, sus múltiples intereses en diferentes empresas así como la gran ejecución de derivados bursátiles que realizan, hacen que **su seguimiento no aporte ningún valor desde el punto de vista analítico**.

Por último, en la ilustración 75, se muestra el resultado de un factor muy importante del análisis en una operación realizada por "*insider"* y es la categorización de la transacción a nivel de propiedad, es decir, si se ha realizado bajo el nombre del "*insider*" llamada "**Direct Holdings (D)**" o a través de un familiar, por ejemplo, esposa o hijos, denominada "***Indirect Holdings* (I)**".

Si se ha realizado bajo el nombre del "*insider"* afecta directamente a su posición financiera y no a acuerdos externos, como fondos familiaries que implican terceras personas. **Si una acción sube o baja afecta directamente a la posición financiera del "*insider*", limitando su capacidad financiera y es por este motivo que el seguimiento de las operaciones de propiedad directa (D) proporciona una mayor de confianza sobre el futuro de la operación**. (Moreland, J. (2000). Profit from Legal Insider Trading. p. 55).

Como se puede observar, **la mayoría de transacciones se reportan afectando directamente a la posición financiera del insider (tipo de propiedad D)** hecho muy positivo para el análisis ya que se automáticamente se descartan todas las transacciones que no lo son.

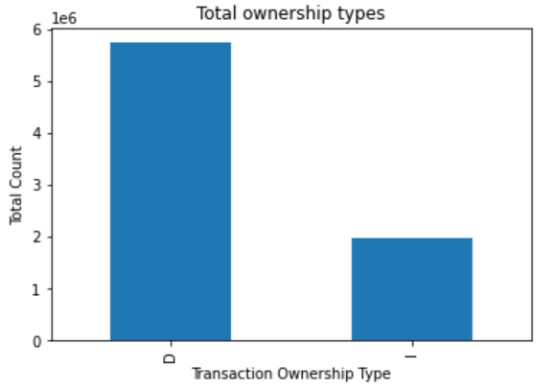


Ilustración 75 Análisis sobre el tipo de propiedad de las operaciones ("direct vs indirect ownership")

5.2.5 CONCLUSIONES DEL ANÁLISIS EXPLORATORIO

Se ha realizado este breve análisis exploratorio con el objetivo de conocer más en profundidad datos sobre como comunican *"insiders"* de las operaciones a través de los documentos *"form 4"*, con la finalidad de investigar su forma de operar y así poder

seleccionar la información que proporcione valor y excluir la que no. (Moreland, J. (2000). Profit from Legal Insider Trading. p. 37).

Como conclusiones podemos afirmar que la información es comunicada sin que se produzcan errores significativos, por lo tanto la información, siempre desde el punto de vista del sistema "EDGAR", es fiable.

También conocemos que los meses de mayor actividad compradora son en el tercer y cuarto trimestre del año, así como el día de la semana de mayor registro de operaciones que es el martes y la hora de estas presentaciones que es a partir de las 16:00 PM, cuando el mercado bursátil ya está cerrado. Es importante destacar una creciente actividad compradora de acciones en el año previo a una recuperación del mercado bursátil.

En lo referente al "*insider"*, suele ser vendedor de acciones y mayoritariamente una persona física, cuyo cargo en la empresa suele ser el de director o con una participación de al menos un 10% en el accionariado y que la mayoría de operaciones se hacen en una propiedad directa (D), que afecta a su posición financiera.

También se ha observado, que se debe ignorar a *"insiders"* con nombres ilustres, como Bill Gates, que para un inversor, en su tarea de seguimiento de operaciones, no aporta ningún valor.

5.3 ANÁLISIS DE LAS OPERACIONES DE COMPRA DE LOS *INSIDERS*

Una vez finalizado el análisis exploratorio, el objetivo es aplicar algunos datos obtenidos así como analizar diferentes comportamientos de los *"insiders"* para poder concluir si un inversor puede obtener algún beneficio del seguimiento de las operaciones de compra realizadas.

5.3.1 ANÁLISIS DEL MERCADO GLOBAL "*DAILY FILING PURCHASES*"

El análisis del mercado global pretende analizar todas las transacciones de compra en conjunto y no de forma individual, con el objetivo de establecer un patrón de comportamiento que se produzca a lo largo del tiempo y que sirva como indicador para establecer una posición de compra en el mercado a largo plazo. Para analizar estas operaciones, se utilizará la vista creada en la base de datos **"daily_purchases_by_directors_and_officers"** que proporciona una lista diaria de las operaciones de compra y venta de los *"insiders"* con rol ejecutivo ("*officers*") durante todo el histórico de datos.

5.3.1.1 FINAL DE LA CRISIS FINANCIERA (2007-2008)

La crisis financiera[13] que se produjo entre los años 2007 y 2008 fue una de las más grandes que se recuerda, perdiendo el mercado bursátil un 54% de su valor y como casi todas las crisis, produjo un mercado bajista que cambio su tendencia en marzo del 2009, produciéndose una nueva tendencia alcista.

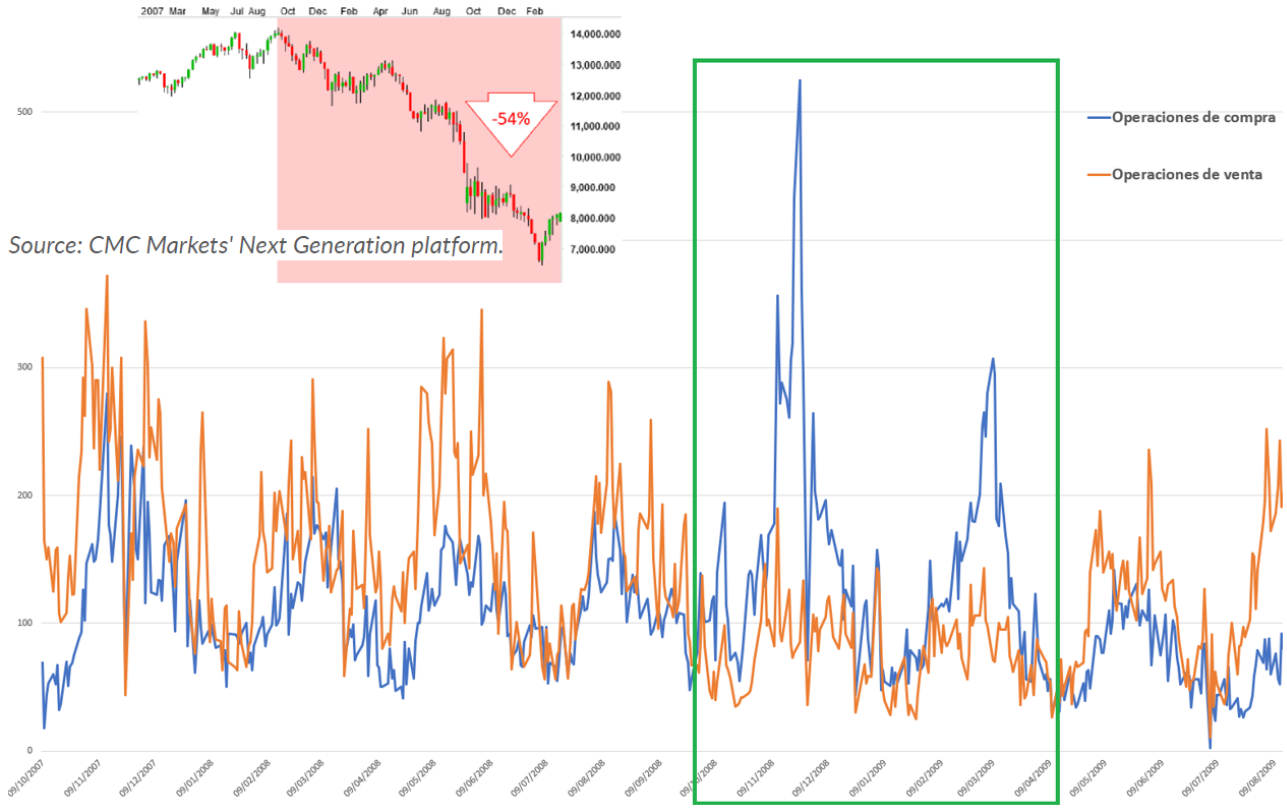


Ilustración 76 Cambio de tendencia diaria en las operaciones de compra tiempo antes de finalizar la crisis financiera.

En la ilustración 76, se puede observar que durante toda la crisis financiera, las operaciones de venta siempre superaron a las de compra pero a partir de septiembre del 2008, coincidiendo con el tercer y cuarto trimestre, la tendencia cambio y las operaciones de compra llegaron a superar en dos y tres veces las operaciones de venta, mostrando los *"insiders",* un importante interés en la compra de acciones de sus empresas. **Estas operaciones se realizaron muchos meses antes de que se iniciara un nuevo mercado alcista**.

De este análisis podemos extraer una nueva conclusión: **los *"insiders"* suelen avanzar un cambio de tendencia con bastante antelación, es decir, suelen comprar bastante tiempo antes de que se produzca un hecho**. Esta conclusión está confirmada. (Moreland, J. (2000). Profit from Legal Insider Trading. p. 71).

[13] Wikipedia. (2020). « *Financial crisis of 2007–2008*». https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008

5.3.1.2 EL "*CRASH*" DEL "*BLACK MONDAY*" (2011)

Detrás de un "*crash*" bursátil siempre existe una noticia relevante que no es la que lo provoca pero sí que resulta ser el detonante. El 8 de agosto de 2011, la deuda soberana de los Estados Unidos sufrió su primera reducción de rating en toda su historia, hecho que provoco el denominado "***Black Monday*** [14]"

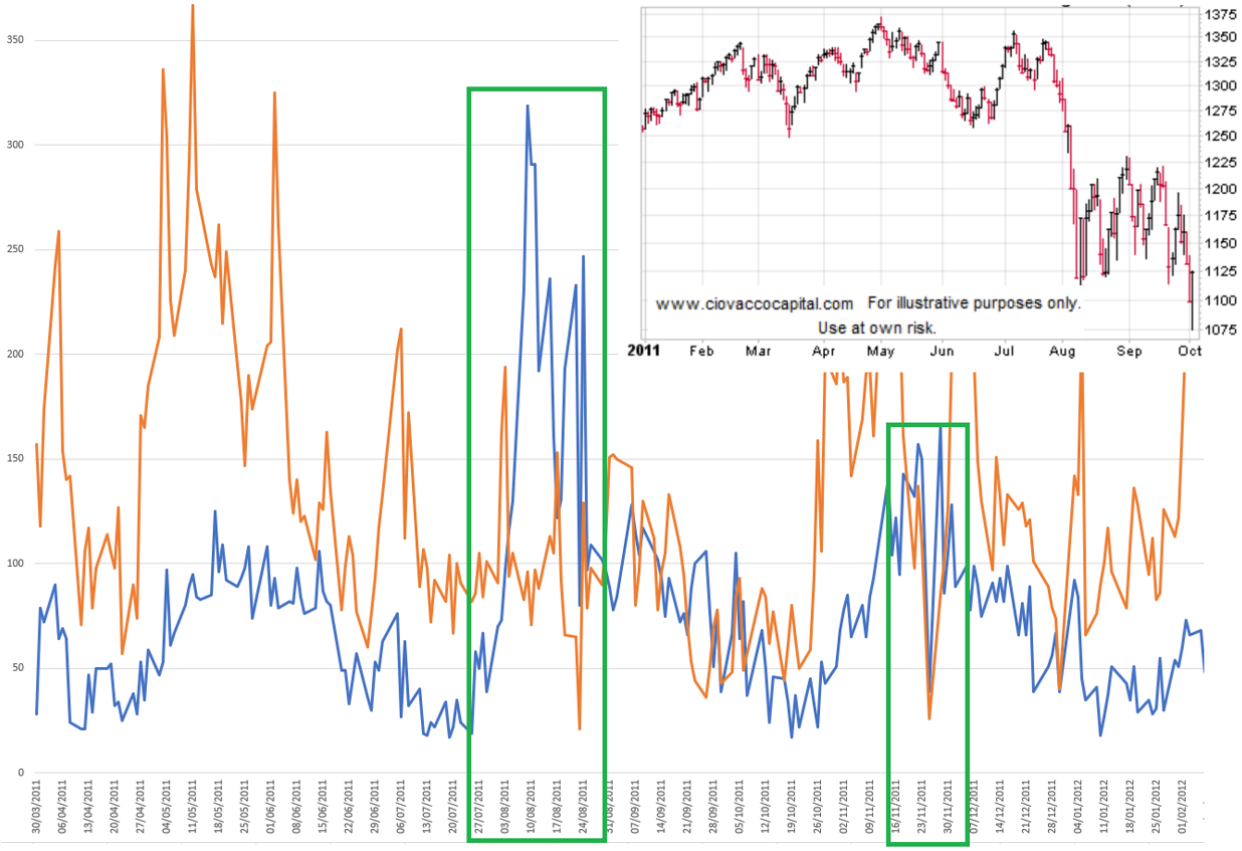


Ilustración 77 Cambio de tendencia diaria en las operaciones de compra tiempo antes y después de producirse el "crash".

En la ilustración 77 podemos ver que antes, durante y tiempo después de que se produjera el "*crash*", los *"insiders"* cambiaron su tendencia de venta de acciones por la de compra y la confirmaron posteriormente. **La conclusión que obtiene es que los *"insiders"* utilizan el pánico bursátil para añadir acciones a su cartera cuando consideran que las acciones de su empresa cotizan con gran descuento** y aunque el mercado siguió a la baja durante muchos meses, volvieron a confirmar su apuesta a mediados de noviembre del 2011, volviendo a incrementar sus posiciones.

Como se ha expuesto anteriormente, los *"insiders"* son el "*Smart Money*" del mercado (Moreland, J. (2000). Profit from Legal Insider Trading. p. 121). Mientras un inversor bursátil vendía sus acciones, los *"insiders"* las acumulaban, lo que indica una firme convicción de que el "*crash*" era únicamente un movimiento corrector del mercado, hecho que se confirmó a partir de enero del 2012, cuando el mercado continuó su tendencia alcista.

[14] Wikipedia. (2020). «*Black Monday (2011)*». https://en.wikipedia.org/wiki/Black_Monday_(2011)

5.3.1.3 EL "*CRASH*" DE LA PANDEMIA DEL COVID-19 (2020)

Al cierre de este proyecto, Europa y el mundo todavía está sufriendo la pandemia del "covid-19" y como es lógico, este tipo de eventos de alcance mundial, generan grandes crisis que se reflejan en la economía y por lo tanto en los mercados bursátiles.

El 20 de febrero del 2020 se produjo el denominado "***Coronavirus Crash*** [15]" que produjo la caída más rápida de un índice bursátil desde el "*crash*" de 1929.

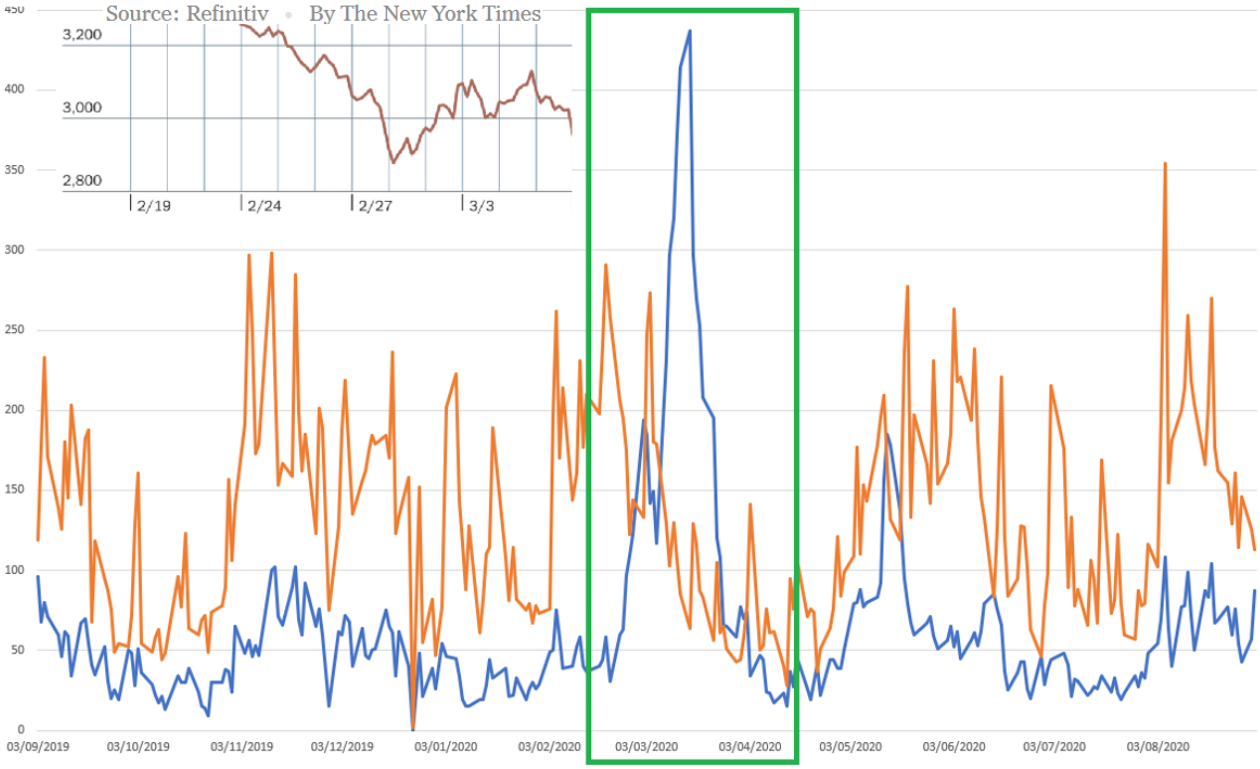


Ilustración 78 Cambio de tendencia diaria en las operaciones de compra después del "Coronavirus Crash".

La ilustración 78 confirma lo expuesto en el anterior punto, el 5.3.1.2, en el que se llega a la conclusión que los *"insiders"*, son oportunistas y por lo tanto aprovechan el momento ideal para comprar acciones, que en muchos casos, coincide con momentos de crisis bursátil. A cierre de este proyecto, a pesar de la grave crisis económica producida por la pandemia, el mercado bursátil ha visto como los índices se revalorizaban en más de un 50% desde el cambio de tendencia de compra por parte de los *"insiders"*.

Como se verá en la siguiente sección, este cambio de tendencia no se produce muy a menudo ya que como se ha comentado, los *"insiders"* son fundamentalmente vendedores de acciones pero si se produce un cambio de tendencia en las operaciones reportadas, es un hecho significativo, que debe analizarse junto con otros indicadores técnicos y económicos.

[15] Wikipedia. (2020). «*Stock Market Crash (2020)*». https://en.wikipedia.org/wiki/2020_stock_market_crash

5.3.1.4 VISIÓN GLOBAL DEL HISTÓRICO DE DATOS (2003-2020)

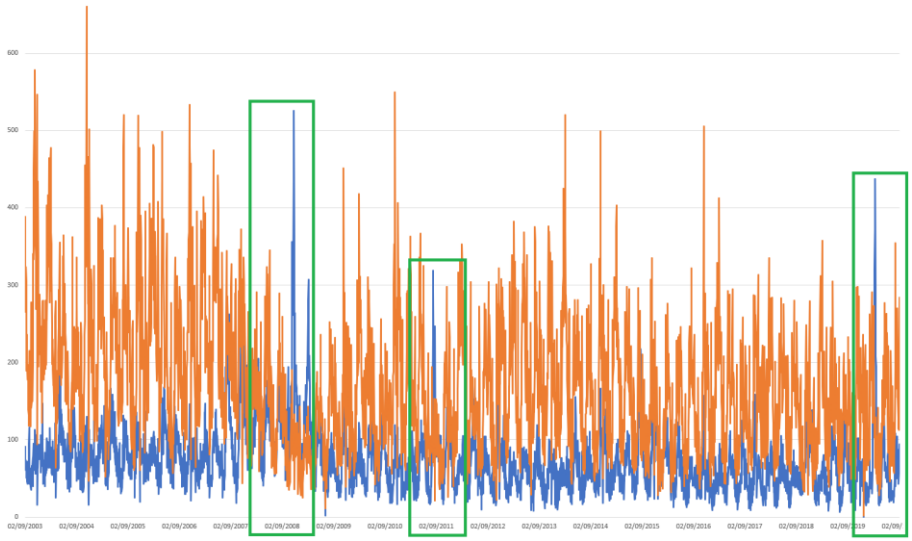


Ilustración 79 Casos en los que se ha producido un cambio de tendencia en las operaciones de compra en el histórico de datos

En la ilustración 79 puede verse las veces que se ha producido un cambio de tendencia en las operaciones de compra desde que la información se ha publicado en formato "XML" en el sistema "EDGAR" (2003-2020). Como se puede observar, este cambio se ha producido tres veces en todo el histórico y ha avanzado futuros movimientos alcistas o a representando confirmaciones de tendencia que cualquier inversor podría haber utilizado para establecer una posición de compra en el mercado.

5.3.2 ANÁLISIS DE COMPAÑÍA COTIZADA

En el análisis de mercado global se ha demostrado que los *"insiders"* suelen avanzarse en sus predicciones, pero siempre desde el punto de vista como grupo inversor colectivo. Se estudiarán diferentes situaciones que pueden producirse y en las que los estos inversores predicen el futuro valor de la acción de una compañía.

5.3.2.1 CONSENSO DE COMPRAS "*BUYING CLUSTER*"

El consenso de compras o "***Buying Cluster***" (Moreland, J. (2000). Profit from Legal Insider Trading. p. 58-60) **sucede cuando dos o más *"insiders"* en un plazo máximo de treinta días realizan compras de acciones de la empresa en la que trabajan o tienen intereses**. Esta compra conjunta, realizada por *"insiders"* con un rol importante dentro de la compañía, como por ejemplo el CEO o el CFO y que afecta a su posición financiera (tipo D) es un hecho significativo, ya que un "*insider"* puede cometer un error al valorar una situación, pero si varios *"insiders"* realizan compras similares en un plazo máximo de treinta días desde la primera compra, es un indicador de que el futuro precio de la acción será superior al precio pagado por el "*insider"*. Se ha desarrollado un método llamado "**process_buying_clusters**" dentro del fichero "**form4/tasks.py**" para la obtención de todos los casos.

En todo el histórico de datos se han producido **154.574** "***Buying Cluster***". Aunque parezca una cifra muy alta, a continuación estableceremos unas reglas para reducir el número total.

La primera regla a aplicar es la exclusión de todos aquellos en que alguno de sus participantes haya realizado la operación de compra en propiedad indirecta (I). Una vez aplicado este filtro, el número total desciende a **65.479**.

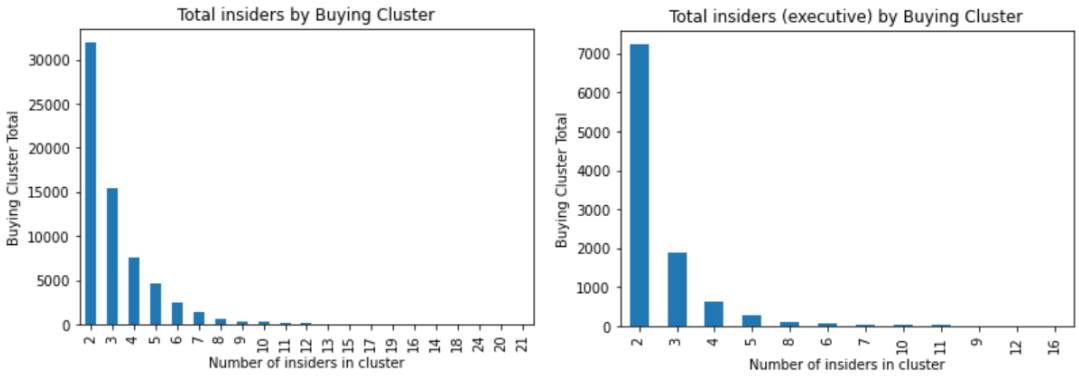


Ilustración 80 Número de insiders por "Buying Cluster" vs "Buying Cluster" formado solo por ejecutivos ("officers")

La ilustración 80 muestra número total de "*insiders*" que han participado en "***Buying Cluster***" y a la derecha únicamente consensos en los que todos los *"insiders"* son ejecutivos (***"officers"***).

Como se puede observar, a medida que el consenso de compras es respaldado por más *"insiders",* el número total de casos disminuye, lo que nos lleva a la siguiente conclusión: **a medida que más *"insiders"* forman parte del consenso, se puede aumentar la fiabilidad, pero los casos disminuyen, es decir la oportunidad se reduce**.

La anterior conclusión obliga a aplicar al total otra regla de exclusión y es la que todos los participantes en el consenso de compras sean ejecutivos, de esta manera excluimos todos aquellos consensos que implican operaciones realizadas por *"insiders"* de menor peso y que pueden implicar operaciones corporativas como la compra de grandes paquetes accionariales. Una vez aplicado el filtro, el número de casos se reduce a **10.327** en el periodo que comprende desde el año 2003 al 2020.

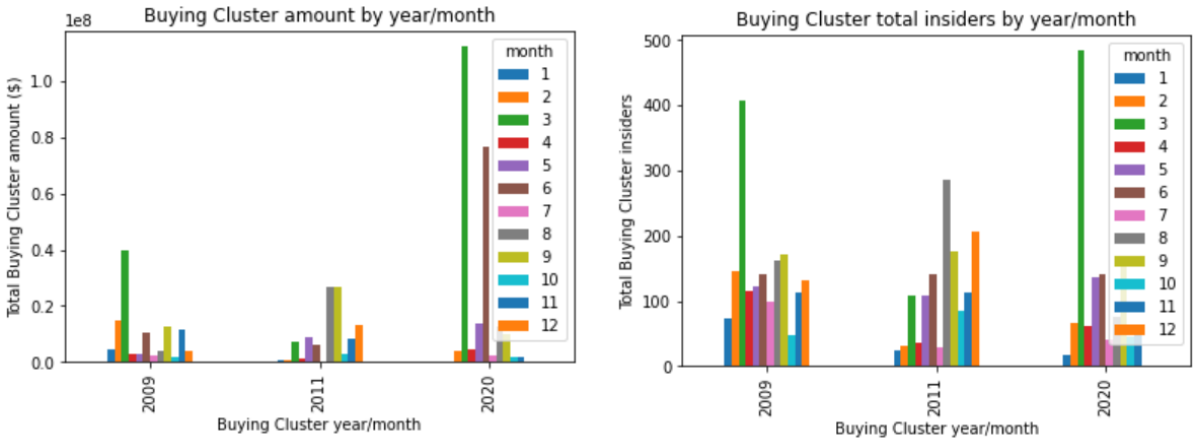


Ilustración 81 "Buying Cluster" en los años de cambio de tendencia del mercado en total ($) y en número de "insiders" participantes

La ilustración 81 muestra un hecho significativo sobre el consenso de compras. Tal y como se verá más adelante, este evento puede analizarse individualmente llevando a cabo un análisis de un "***Buying Cluster***" en concreto, o como herramienta para confirmar sucesos que puedan producirse en el mercado bursátil. En los años en los que se ha producido un cambio de tendencia en el mercado, durante los meses previos, hubo una gran participación de *"insiders"* en operaciones de compra, produciendose un gran aumento del total en dólares en las operaciones hechas en conjunto. Esta hecho, confirma todo lo expuesto en la sección de "**análisis de mercado**" y permite a un inversor añadir más fiabilidad al indicio de confirmación o cambio de tendencia en el mercado.

En el caso de producirse un suelo de mercado, el número de empresas en las que se producen estas operaciones de "***Buying Cluster***" aumenta el año anterior del cambio de tendencia, como podemos ver en la ilustración 82 en los datos del año 2008 en el que 1.031 empresas fueron objetivo de consensos de compras.

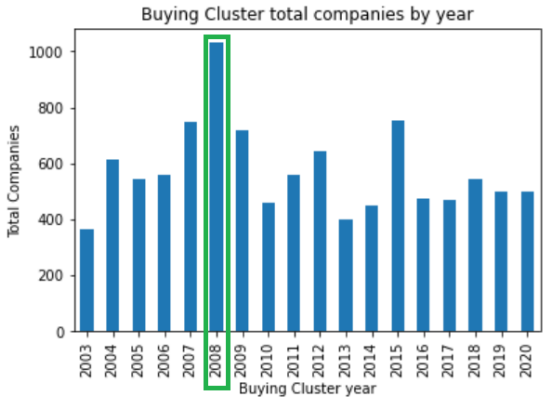


Ilustración 82 Número de compañías en operaciones de "Buying Cluster" en todo el histórico de datos

El consenso de compras, como se ha visto, permite escoger el mejor momento para establecer una posición de ventaja en el mercado y aunque esto podría ser suficiente premio para un inversor, existe otra posibilidad que puede aportar un beneficio mayor.

Un mayor retorno de la inversión se podría producir a través del análisis de un "***Buying Cluster***" concrecto, con el objetivo de posicionarse en el accionariado, adquiriendo acciones de la empresa, a un precio similar al que pagaron los *"insiders"* que participaron en consenso, esperando una revalorización de las acciones en el plazo mínimo de un año. **En algunos casos estas revalorizaciones pueden llegar a ser de un 100% o más**.

No todos los "*Buying Cluster*" finalizan de forma correcta, es decir, produciendo una revalorización al alza del precio por acción, pero como se ha visto, si el número de participantes en el consenso es alto, existe más fiabilidad. Como se verá, la mayoría de estas operaciones en conjunto estan asociadas a eventos futuros que se producirán en la empresa, como un un aumento de beneficios en las cuentas trimestrales o la adquisición de la empresa por parte de un tercero.

Se van a analizar 3 casos significativos que hubieran producido un gran rentabilidad para un inversor que hubiera participado en el consenso.

5.3.2.1.1 Caso MeadWestvaco Corporation (NYSE: MWV)

Las acciones de MeadWestvaco Corporation[16] empresa dedicada al sector de la paquetería industrial, durante la crisis financiera (2007-2008), habían sufrido un gran castigo bursátil, pasando de un máximo de 31 dólares por acción a un mínimo de 14.

Durante el intervalo del 31-10-2008 al 19-11-2008 se produjeron las siguientes compras por parte de los *"insiders"* de la compañía, detalladas a continuación:

Insider	Fecha	Rol	Acciones	Precio/acción	Total $	AN documento *"form 4"*
Willkie Wendell L II	31-10-2008	Vice presidente Senior	4.000	14,22	56.887	0001159297-08-000133
Rajkowski E Mark	05-11-2008	Director financiero (CFO)	6.000	14,73	88.380	0001159297-08-000144
Schreiner Linda V	12-11-2008	Vice presidente Senior	290	12,19	3.535	0001159297-08-000145
Luke John A Jr	12-11-2008	Presidente (CEO)	10.000	11,52	115.200	0001159297-08-000146
Watkins Mark T	19-11-2008	Vice presidente Senior	2.000	11,3169	22.633	0001159297-08-000147

La información que contiene la tabla expone claramente el concepto de un "***Buying Cluster***".

Como se puede observar, en un periodo máximo de treinta días, cinco "*insiders*" que ejercían roles ejecutivos dentro de la compañía, compraron un total de 22.290 acciones a un precio medio de 12,79 dólares. Se observa que las compras, al producirse en diferentes fechas, son a un precio distinto. Este dato, es positivo, ya que se puede afirmar que no son compras programadas, es decir, se elimina la posibilidad de que no sea un verdadero "***Buying Cluster***" " (Moreland, J. (2000). Profit from Legal Insider Trading. p. 60)

Ilustración 83 "Buying Cluster" en MeadWestvaco Corp.

Un inversor que hubiera utilizado esta información como confirmación para un posicionamiento en el mercado, en fechas posteriores a la última compra del "*insider*", podría haber adquirido acciones dentro del mínimo histórico de 5 años: 8,20 dólares por acción.

Como se ha visto anteriormente, los *"insiders"* suelen avanzar movimientos, pero al comprar con antelación, la compra no suele ser en el mínimo histórico sino cuando ellos consideran que el precio de la acción cotiza con descuento, ya que en este caso, el precio por acción, se redujo en un gran porcentaje antes de cambiar a una tendencia alcista tal y como se puede observar en la ilustración 80. La acción se revalorizo desde los 8,20 hasta los 54 dólares, precio por el que la empresa fue adquirida.

[16] Wikipedia. (2020). «*MeadWestvaco Corporation*». https://en.wikipedia.org/wiki/MeadWestvaco

5.3.2.1.2 Caso Talen Energy Corporation (NYSE:TLN)

Dado que el anterior caso se produjo durante una crisis financiera, se podría pensar que estas compras consensuadas de acciones únicamente se producen cuando hay alguna crisis en el mercado bursátil y por lo tanto, las acciones, cotizan siempre a precios muy bajos. Como se mostrará, esto no siempre es así, ya que este tipo de situaciones pueden producirse en cualquier tipo de mercado y sobre todo en eventos corporativos complejos.

Talen Energy Corporation[17] se fundó en el año 2015 como resultado del "***spin-off***" o separación de la unidad de negocio energético de la compañía madre PPL. A esta separación se le unió una fusión corporativa para formar una nueva empresa que cotizara en el mercado bursátil. Este tipo de eventos corporativos son complejos y en muchas ocasiones requieren la aprobación de organismos reguladores. El 23-04-2015 se recibió la aprobación y el 18-05-2015 las acciones de la nueva empresa empezaron a cotizar en el NYSE a un precio de 27 dólares.

En los siguientes meses, las acciones de la compañía perdieron mucho valor, fundamentalmente por la posibilidad de litigios judiciales[18] y llegaron hasta mínimos de 6 dólares por acción a finales de diciembre de 2015.

En el intervalo del 09-12-2015 al 23-12-2015 se produjo el siguiente "***Buying Cluster***" que a continuación se detalla:

Insider	Fecha	Rol	Acciones	Precio/acción	Total $	AN documento *"form 4"*
Farr Paul A	09-12-2015	Presidente (CEO)	10.000	6,7570	67.570	0001159297-08-000133
Schinski James E	09-12-2015	Vice presidente Senior	30.816	6,7540	208.131	0001209191-15-084621
Rausch Timothy S	17-12-2015	Chief Nuclear Officer	12.100	5,9180	71.607	0001209191-15-086340
Mcguire Jeremy R	21-12-2015	Director Financiero (CFO)	10.000	6,4100	64.100	0001209191-15-086813

En total, los *"insiders"* adquirieron 62.916 acciones a un precio medio de 6,45 dólares por acción. Como se puede ver en la ilustración 81, el precio de la acción cambió a una tendencia alcista poco tiempo después, llegando hasta los 14 dólares por acción, precio por el que fue adquirida la compañía.

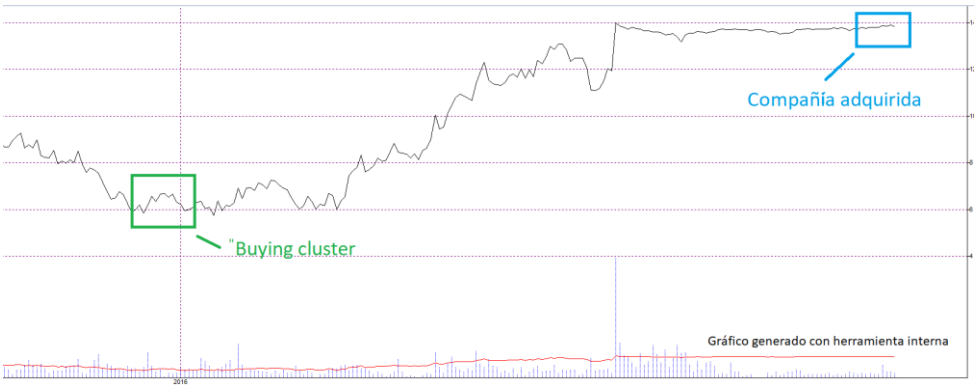


Ilustración 84 "Buying Cluster" en Talen Energy Corp.

[17] Wikipedia. (2020). «*Talen Energy*». https://en.wikipedia.org/wiki/Talen_Energy

[18] Bizjournals. (2015). «*Another competitor says it may sue to stop FirstEnergy's PPA*». https://www.bizjournals.com/columbus/blog/ohio-energy-inc/2015/12/another-competitor-says-it-may-sue-to-stop.html

5.3.2.1.3 Caso Roanoke Electric Steel Corp. (NASDAQ:RESC)

Los *"insiders"*, no sólo son capaces de reconocer cuando la acción de su empresa cotiza con descuento sino que al estar al frente del día a día en las operaciones pueden avanzarse a futuros eventos, como por ejemplo, un cambio positivo en el resultado operativo de un trimestre.

Las acciones de Roanoke Electriz Steel Corp.[19]. cotizaban a 6 dólares por acción a mediados de mayo del 2003. El mercado había estado en recesión durante dos años y los beneficios trimestrales de empresas como Roanoke, habían sido negativos.

En el intervalo del 23-05-2003 al 13-06-2003 se produjo el siguiente "*Buying Cluster*" que a continuación se detalla:

Insider	Fecha	Rol	Acciones	Precio/acción	Total $	AN documento *"form 4"*
Smith Donald G	23-05-2003	Presidente (CEO)	4.000	6,17	24.680	0000084278-03-000008 0001229452-04-000006
Cartledge George B Jr	23-05-2003	Director	8.000	6,2575	50.006	0000084278-03-000009
Cartledge George B Jr	27-05-2003	Director	2.000	6,39	12.780	0001229459-03-000001
Logan George W	02-06-2003	Director	10.000	7,5115	75.115	0001013321-03-000001
Higgins Donald R	12/13-06-2003	Vice presidente	103	6,90	711	0001229455-03-000001

La empresa, el 15-03-2004, presento los resultados[20] del primer trimestre del 2004 que por primera vez en dos años arrojaron beneficios.

Las acciones de la empresa se revalorización desde un mínimo de 5,80 dólares por acción hasta los 33,28, precio por el que fue adquirida.

Ilustración 85 "Buying Cluster" en Roanoke Electric Steel Corp.

[19] Wikipedia. (2020). « *ROANOKE ELECTRIC STEEL CORPORATION FIRST QUARTER 2004 RESULTS*». https://sec.report/Document/0000084278-04-000002/

[20] Wikipedia. (2020). « *ROANOKE ELECTRIC STEEL CORPORATION FIRST QUARTER 2004 RESULTS*». https://sec.report/Document/0000084278-04-000002/

5.3.2.2 COMPRA INTELIGENTE "*IQ BUY*"

La compra inteligente o "***IQ Buy***" no es un hecho que suceda en el tiempo, sino que es **la categorización de un *insider* dentro de un grupo de inversores inteligentes que han realizado dos o más operaciones de compra con una rentabilidad positiva mínima a un año**. Se excluyen todas aquellas operaciones de compra en el que el precio de la acción sea inferior a cinco dólares. De esta forma, se ignoran todas aquellas operaciones que no serían de interés para un inversor, por no estar aconsejada su adquisición[21].

En esta sección se exponen tres casos que no sólo cumplen los criterios de compra inteligente, sino que se ha exigido que las compras se hayan realizado en compañías diferentes, algo que ocurre en muy pocas ocasiones.

Un "*insider"* puede predecir el movimiento al alza de una acción una o más veces dentro de su empresa, pero si este hecho ocurre en diferentes empresas, es candidato a ser controlado por su extrema fiabilidad.

Se ha desarrollado un "*script*" llamado "**iq.py**" dentro de la carpeta del software "ETL" para obtener los resultados. Queda fuera del alcance de este proyecto, almacenar el resultado de las operaciones con el fin de establecer un "*ranking*" para categorizar a este grupo.

Ejecutando el "*script*" sobre el *"data warehouse"* se han obtenido los siguientes resultados:

- **216.538** casos con rentabilidad positiva vs **214.506** con rentabilidad negativa a un precio por acción mayor o igual cinco dólares.
- **8.005** casos de con operaciones de compra en la misma empresa con rentabilidad mínima positiva a un año de un 20% y precio por acción mayor o igual cinco dólares**.**
- **244** casos con operaciones de compra en un máximo de dos empresas con rentabilidad positiva mínima a un año de un 20% y precio por acción mayor o igual cinco dólares**.**
- **22** casos con operaciones de compra en tres o más empresas con rentabilidad positiva mínima a un año de un 20% y precio por acción mayor o igual cinco dólares.
- La rentabilidad media positiva de todos los casos ha sido de un **34,64%.**

Una vez conocidos todos estos datos, una futura línea de trabajo para el proyecto sería la categorización de los *"insiders"* en diferentes grupos y a través de un sistema

[21] The Motley Fool. (2020). « Why Buying Cheap Stocks Is the Wrong Investing Strategy». https://www.fool.com/investing/2020/10/17/why-buying-cheap-stocks-is-the-wrong-investing-str/?source=awin&awc=12195_1605879211_4bd03ee3537e8ab64f0a7e1d204adcde&utm_source=aw&utm_medium=affiliate&utm_campaign=101248

de alertas comunicar al inversor cuando se produce una operación cumpliendo los criterios anteriormente especificados.

Como se puede observar, los *"insiders"* que toman decisiones en diferentes empresas de forma satisfactoria son muy pocos.

5.3.2.2.1 Caso Glenn R. Simmons (CIK: 0001188357)

Es muy usual que los *"insiders"* formen parte de consejos de administración en diferentes empresas.

Glenn R Simmons, en el histórico de datos, realizó las siguientes operaciones de compra en diferentes empresas en las que ejercía el rol de presidente del consejo, obteniendo los siguientes resultados:

Fecha	Empresa	Rol	Acciones	Precio/compra	Fecha 1 año	Precio 1 año	Rev. % 1 año	AN doc. "form 4"
14-01-2004	COMPX International Inc (NYSE:CIX)	Presidente consejo	200	6,71	14-01-2005	16,45	+136,69 %	0001049606-04-000003
24-02-2009	Titanium Metals Corp. (NYSE: TIE)	Presidente consejo	13.400	6,09	24-02-2010	11,65	+90,98 %	0001011657-09-000003
08-07-2009	VALHI, Inc (NYSE: VHI)	Presidente consejo	5.000	6,40	08-07-2010	14,38	+122,22 %	0000059255-09-000131

La tabla anterior refleja las 3 operaciones con mayor rentabilidad a un año de Glenn R Simmons, mostrando que es un "*insider*" que suele operar en el mercado cuando existen descuentos importantes en el precio de las acciones de las compañías en las que participa. Se pueden observar en algunos casos rentabilidades de más de un 100%.

5.3.2.2.2 Caso Michael Jennings (CIK: 0001319731)

Los *"insiders"* también pueden ejercer cargos en múltiples empresas que no sean únicamente la presidencia de consejos sino roles de dirección directa como el de CEO.

Michael Jennings, en el histórico de datos, realizó las siguientes operaciones de compra en diferentes empresas en las que ejercía el rol de presidente del consejo y CEO, obteniendo los siguientes resultados:

Fecha	Empresa	Rol	Acciones	Precio/compra	Fecha 1 año	Precio 1 año	Rev. % 1 año	AN doc. "form 4"
26-02-2006	Holly Energy Partners LP (NYSE: HEP)	CEO	2.936	29,14	27-02-2007	35,83	+23,64 %	0000059255-09-000131
09-09-2010	Frontier Oil Corp. (NYSE:FTO)	Presidente consejo/CEO	5.000	12,45	30-06-2011	32,31	+159,85 %	0001140361-10-036662
21-11-2011	HollyFrontier Corp. (NYSE: HFC)	Presidente / CEO	2.000	24,10	21-11-2011	41,40	+102,18 %	0001209191-11-057372

En la tabla anterior, podemos ver que las rentabilidades son muy similares al anterior caso estudiado pero se añade un factor diferencial importante, ya que una de las operaciones, a su vez, forma parte de un "***Buying Cluster***" y por lo tanto ofrece un grado aún mayor de fiabilidad. **Una operación que forma parte de varios sucesos tiene mayor credibilidad para un inversor**.

5.3.2.2.3 Caso Richard D. Crowley Jr (CIK: 0001184854)

Como se verá en la siguiente sección, los roles ejecutivos más importantes como el CEO (presidente) y el CFO (director financiero) añaden más credibilidad a una compra inteligente. En este caso el Richard D Crowley Jr. Ejercía como director financiero en las tres empresas en las que realizó las compras y por sus funciones conocía perfectamente el estado de las finanzas en estas empresas.

Este "*insider*", en el histórico de datos, realizó las siguientes operaciones de compra en diferentes empresas, obteniendo los siguientes resultados:

Fecha	Empresa	Rol	Acciones	Precio/compra	Fecha 1 año	Precio 1 año	Rev. % 1 año	AN doc. "form 4"
26-02-2006	Intersil Corp. (NASDAQ: ISIL)	Director Financiero	1.500	11,60	23-02-2007	22,49	+93,71 %	0001140361-16-063409
06-02-2008	MICREL INC (NASDAQ:MCRL)	Director Financiero	1.000	6,09	06-02-2009	7,92	+33,22 %	0001184854-08-000001
08-06-2010	Integrated Device Technology Inc (NASDAQ: IDTI)	Director Financiero	500	5,15	08-06-2012	7,57	+43,30 %	0001230186-10-000030

Como conclusión se puede afirmar que **el seguimiento de las compras inteligentes puede proporcionar al inversor un indicador fiable de no sólo cuando debe comprar, sino también a qué precio**, ya que es muy importante no comprar acciones a un precio muy distinto al que las adquirió el *insider*. (Moreland, J. (2000). Profit from Legal Insider Trading. p. 73)

5.3.2.3 COMPRA MAYOR DE ALTA DIRECCIÓN "*CEO/CFO 100K BUY*"

La compra mayor de alta dirección "*CEO/CFO 100k buy*" **es el seguimiento de todas las operaciones de compra con un importe total igual o superior a 100.000 dólares, en propiedad directa (D) y realizadas por uno de los *"insiders"* con más peso dentro de la empresa: el CEO (presidente) o el CFO (director financiero)**.

De la misma forma que en anteriores análisis, este evento también puede confirmar datos a nivel de "*análisis de mercado*" como la continuación o el inicio de una nueva tendencia alcista.

Se ha definido una vista en la base de datos llamada "**ceo_cfo_buy_100_k**" que se utilizará para la realización de este análisis.

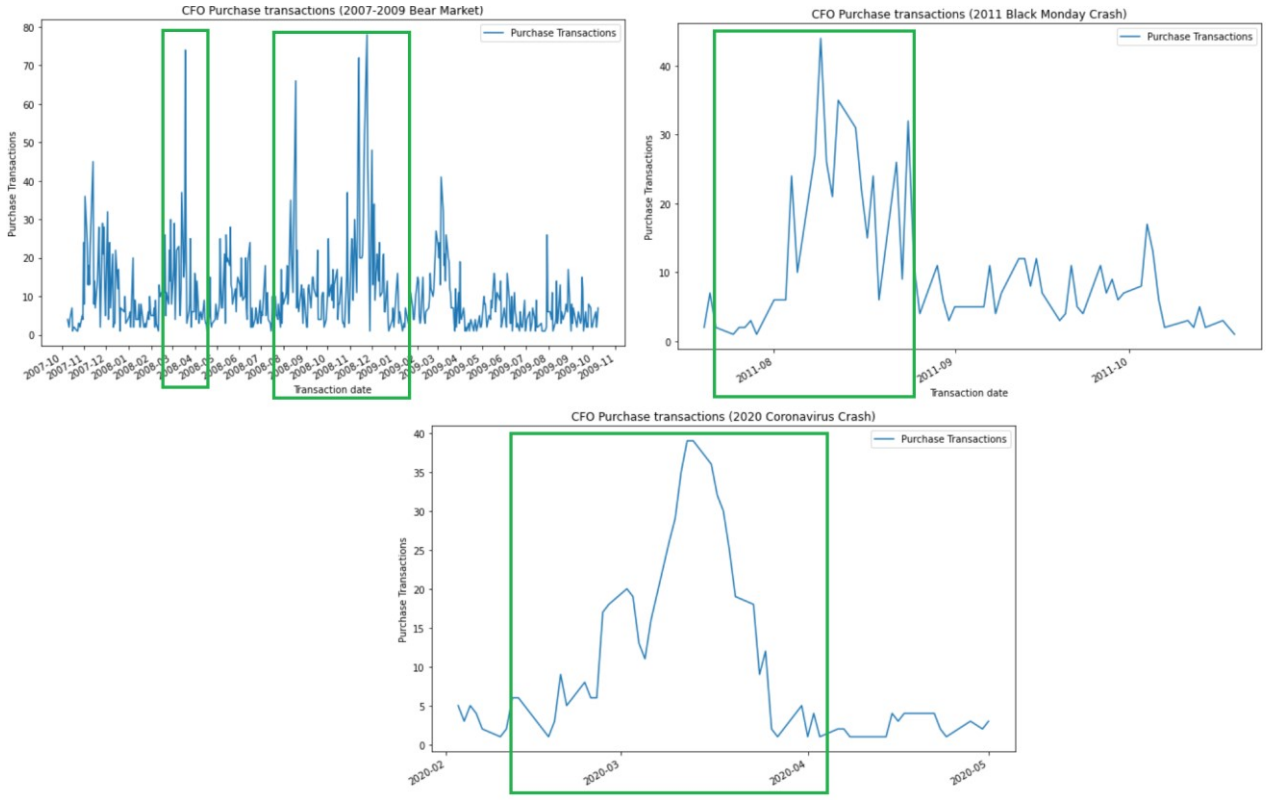


Ilustración 86 Cambio de tendencia en las operaciones de compra de los insiders con rol de CFO (director financiero) durante las crisis bursátiles

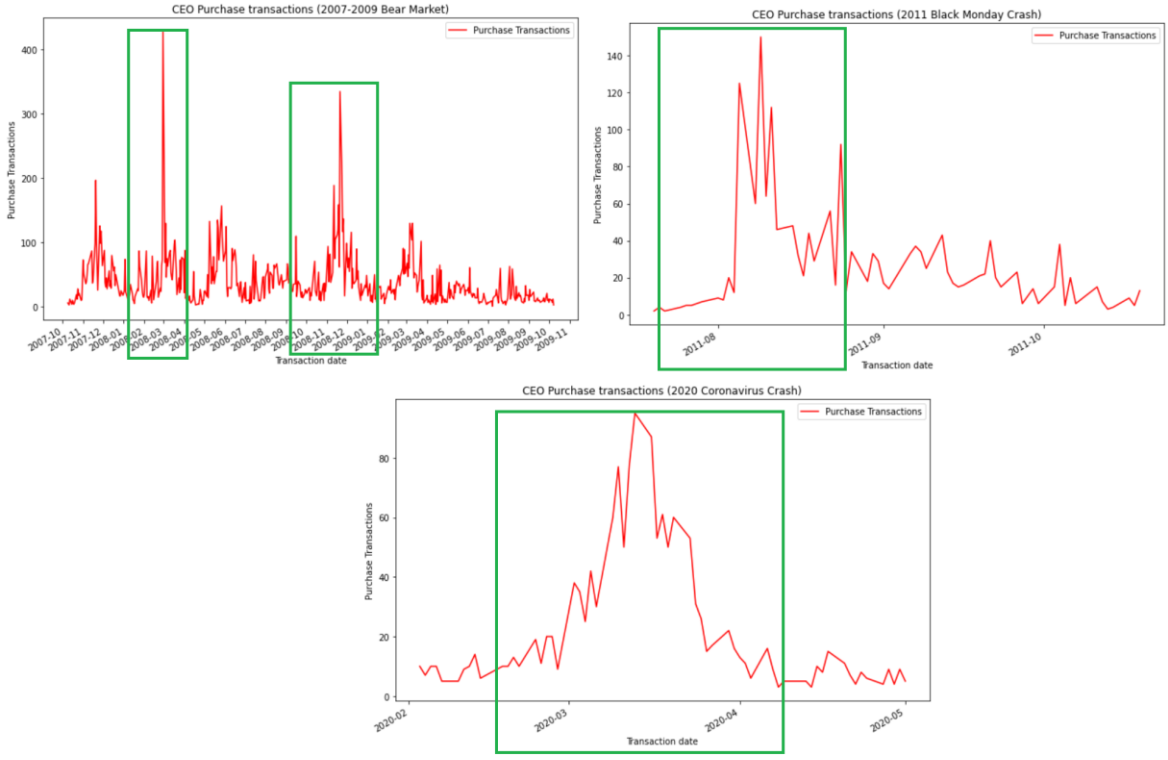


Ilustración 87 Cambio de tendencia en las operaciones de compra de los insiders con rol de CEO (presidente) durante las crisis bursátiles

Las ilustraciones 86 y 87, muestran el cambio de tendencia en las operaciones de compra de los *"insiders"* con roles de CEO y CFO durante las diferentes crisis de mercado, exploradas en la sección de "***análisis de mercado***". Ambos casos han experimentado un patrón de comportamiento muy similar. **Los *"insiders" con cargos de CEO/CFO* confirman la continuación o el inicio de una tendencia de precios al alza**.

En el histórico de datos (2003-2020) para las compras de alta dirección "*CEO/CFO 100k buy"* se han obtenido los siguientes resultados:

- 2.899 operaciones de compra de alta dirección con rentabilidad mínima positiva a un año de un 20% y precio por acción mayor o igual cinco dólares.
- 2.451 casos han sido realizados por *"insiders"* con rol de CEO (presidente)
- 448 casos han sido realizados por *"insiders"* con rol de CFO (director financiero)
- La rentabilidad positiva media a un año para los CEO ha sido del 57,92 %
- La rentabilidad positiva media a un año para los CFO ha sido del 68,23 %

A continuación se analizan 3 casos significativos.

Fecha	Rol	Acciones	Precio/compra	Fecha 1 año	Precio 1 año	Rev. % 1 año	AN doc. "form 4"
21-02-2016	Director Financiero	10.000	10,69	23-01-2017	21,89	+101,75 %	0000723125-16-000150

5.3.2.3.1 Caso Micron Technology Inc. y Ernest E. Maddock

Ernest E. Maddock, CFO de la compañía Micron Technology Inc. (NASDAQ: MU), adquirió un paquete importante de acciones en febrero de 2016 justo antes de que finalizara la gran caída de valor (-60%) que habían sufrido las acciones de la compañía[22] durante el año 2015.

En la ilustración 88, una vez más, se puede observar un movimiento inteligente del "*Smart Money*", conocedor de una información[23] que el resto de inversores desconocen.

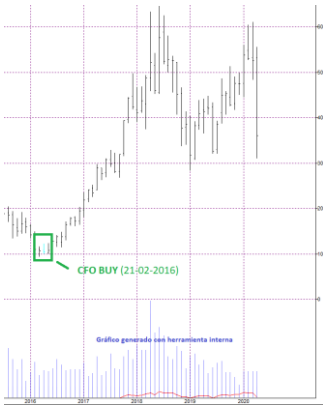


Ilustración 88 Compra de acciones del CFO de la empresa anunciando un cambio de tendencia del precio al alza (10,69 a 60 dólares)

[22] The Motley Fool. (2020). « Why Micron Technology, Inc. Fell 60% in 2015». https://www.fool.com/investing/general/2016/01/05/why-micron-technology-inc-fell-60-in-2015.aspx

[23] The Motley Fool. (2020). « Why Micron Technology, Inc. Shares Rose 88% in 2017». https://www.fool.com/investing/2018/01/05/why-micron-technology-inc-shares-rose-88-in-2017.aspx

En este caso, la adquisición de una cantidad tan alta de acciones y su cargo dentro de la empresa, anunciaba a los inversores una gran oportunidad de compra.

Las acciones fueron adquiridas a 10,69 y a principios de 2018 cotizaban a casi 60 dólares.

5.3.2.3.2 Caso CREE Inc. y Charles M. Swoboda

Fecha	Rol	Acciones	Precio/compra	Fecha 1 año	Precio 1 año	Rev. % 1 año	AN doc. "form 4"
05-12-2008	Presidente (CEO)	10.000	13.05	07-12-2009	51.18	+289,79 %	0000895419-08-000074

Charles M. Swoboda, CEO de la empresa CREE Inc.[24] (NASDAQ: CREE), en diciembre de 2008 adquirió 10.000 acciones, en un momento en el que la crisis financiera llegaba a su fin. Este caso confirma el análisis inicial de esta sección, en el que se establecía una relación directa entre los cambios o confirmación de una tendencia en el mercado y las grandes adquisiciones de acciones por parte de *"insiders"* con un peso específico dentro de la compañía.

La ilustración 89, muestra la adquisición de un paquete accionarial por parte del CEO y su posterior revalorización. También se puede apreciar el poco volumen de operaciones de compra que existe en el momento que los "*insiders*" deciden realizar sus operaciones y en fechas posteriores.

Este poco volumen, demuestra que los inversores no replican el movimiento de compra del "*insider*" y que operan una vez se han producido grandes revalorizaciones de precio, perdiendo una gran oportunidad de inversión.

Las acciones fueron adquiridas a 13,05 dólares y un año después cotizaban a 51,18. Como vemos en la ilustración, las acciones llegaron a un máximo 85 dólares.

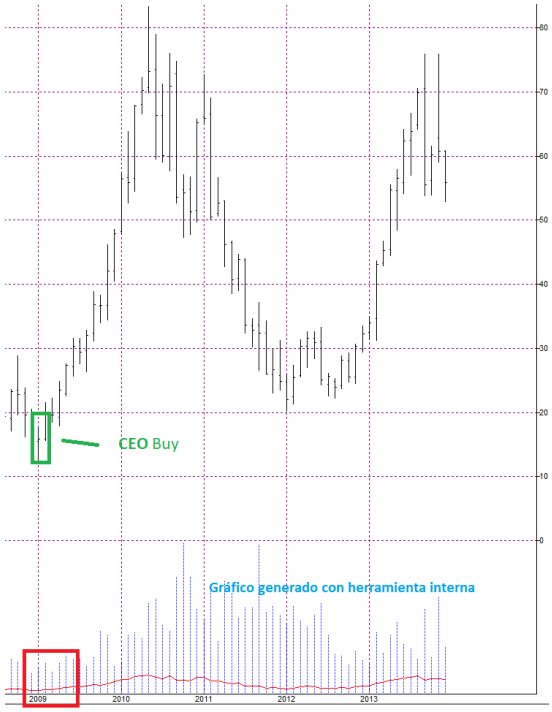


Ilustración 89 Compra de acciones del CEO de la empresa anunciando un cambio de tendencia del precio al alza (13,05 a 80 dólares)

[24] Wikpedia (2020). « Cree Inc.». https://en.wikipedia.org/wiki/Cree_Inc.

5.3.2.3.3 Caso Jazz Pharmaceuticals Inc. y Kathryn E. Falberg

Fecha	Rol	Acciones	Precio/compra	Fecha 1 año	Precio 1 año	Rev. % 1 año	AN doc. "form 4"
30-08-2010	Director Financiero	25.689	8,53	30-08-2011	42,50	+400 %	0001232524-10-000021

Kathryn E. Falberg, CFO de la empresa Jazz Pharmaceuticals Inc. adquirió un gran paquete de acciones, en un momento en el que la empresa cotizaba en el mínimo histórico. Las empresas farmacéuticas son un caso especial dentro del mercado bursátil ya que la cotización de sus acciones se basa, en muchas ocasiones, no sólo en sus beneficios, sino en la aprobación de sus medicamentos en desarrollo por las correspondientes agencias gubernamentales[25].

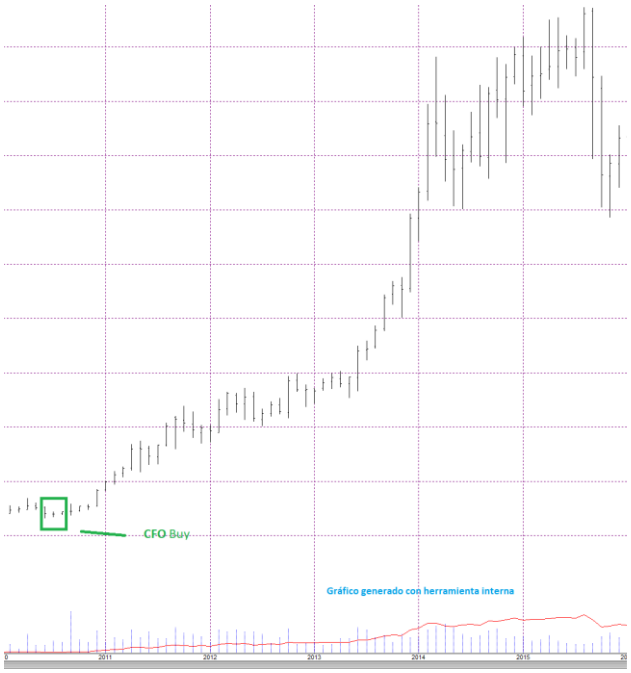


Ilustración 90 Compra de acciones del CFO de la empresa anunciando un cambio de tendencia del precio al alza (8,53 a 180 dólares)

Este caso de estudio, no difiere en mucho a los anteriores estudiados, el oportunismo, una vez más, en la compra de unas acciones en un momento de cotización baja, es algo recurrente que se ha podido observar en todo el análisis que se ha realizado.

La ilustración 90 muestra la evolución de las acciones de la empresa farmacéutica durante los años posteriores a la compra del "*insider*" y cómo se puede observar obtuvo unos grandes beneficios. Las acciones se revalorizaron desde los 8,53 a más de 180 dólares por acción en los años posteriores a la adquisición del "*insider*".

De los 2.899 casos que se han producido en el histórico, se ha elegido este en último lugar porque es un claro ejemplo de que en la toma de una decisión de inversión, hay que analizar profundamente toda la información que aparece en el documento *"form 4"*, ya que podría incluir algún tipo de nota que restara valor a la operación y que sólo la intervención humana puede determinar si es una operación a descartar.

```
<footnotes>
  <footnote id="F1">The price reported in Column 4 is a weighted average
  price. These shares were purchased in multiple transactions at prices
  ranging from $8.48 to $8.58, inclusive.</footnote>
  <footnote id="F2">The price reported in Column 4 is a weighted average
  price. These shares were purchased in multiple transactions at prices
  ranging from $8.85 to $9.10, inclusive.</footnote>
  <footnote id="F3">The reporting person undertakes to provide to Jazz
  Pharmaceuticals, Inc., any security holder of Jazz Pharmaceuticals, Inc.,
  or the staff of the Securities and Exchange Commission, upon request,
  full information regarding the number of shares acquired at each separate
  price within the ranges set forth in footnotes (1) and (2) to this Form
  4.</footnote>
</footnotes>
```

Ilustración 91 Notas añadidas por el insider para informar de detalles adicionales de la operación de compra

En el *"data warehouse"*, se han incluido todas las notas que los *"insiders"* añaden a las

[25] First World Pharma. (2020). « Jazz Pharmaceuticals shares rise on Barclays boost». https://www.firstwordpharma.com/node/666723?tsid=17

operaciones de compra y que pueden condicionar la decisión de una inversión.

En la ilustración 91, podemos ver que el Kathryn E. Falberg, informó a través de notas añadidas en el documento *"form 4"*, que había adquirido más acciones a precios distintos y que la media de todas sus adquisiciones era de 8,54 dólares por acción. Este dato, no cambia el resultado final del análisis de este caso, ya que la operación para un inversor hubiera sido excelente, pero muestra que es necesario analizar información adicional para determinar si en la operación existe algún dato que permita descartarla. (Moreland, J. (2000). Profit from Legal Insider Trading. p. 70)

5.4 CONCLUSIÓN FINAL Y FUTURAS LÍNEAS DE TRABAJO

Un inversor puede utilizar el seguimiento de las operaciones de compra realizadas por los *"insiders"* como un indicador adicional para la toma de una decisión de inversión. Es importante recordar que cualquier decisión de inversión debe ser reforzada con múltiples indicadores y el uso de este seguimiento es sólo uno de ellos.

Como se ha podido demostrar, los *"insiders"* son fundamentalmente vendedores de acciones pero en el momento que realizan una operación de compra, se produce un hecho significativo que al ser analizado puede proporcionar una información muy valiosa sobre el futuro precio de una acción o la tendencia de un mercado bursátil.

Los *"insiders"* suelen ser personas físicas que reportan mayoritariamente las operaciones a través de los documentos *"form 4"* al cierre del mercado y las operaciones de compra se aceleran en el tercer y cuarto trimestre del año.

En momentos de crisis bursátil, los *"insiders"* suelen avanzar los cambios de tendencia con bastante antelación indicando un próximo suelo de mercado.

Dentro de sus compañías, los *"insiders"* suelen apostar por las acciones de sus empresas en momentos de crisis, ya sea en un momento de delicado para la compañía en la que tienen intereses o en un mercado bursátil a la baja, formando eventos como el "*Buying Cluster*" en el que varios de ellos participan con un objetivo común: obtener un beneficio económico.

El *"data warehouse"* desarrollado ha permitido también extraer información que proporciona información global del mercado y que ofrece la posibilidad de categorizar los *"insiders"* en grupos como "*IQ buy*" o la compra mayor de alta dirección *"CEO/CFO 100K BUY*" y cuyo objetivo final es el aumento de la fiabilidad para que una operación de compra realizada por un *insider,* se convierta en una gran oportunidad para un inversor.

Una futura línea de trabajo sería la categorización de los *"insiders"* en grupos de inversión y así poder establecer un "*ranking"* que proporcionara información sobre el éxito o fracaso en sus operaciones. Para ello sería necesario la integración en el *"data warehouse"* de un sistema de cotizaciones de acciones que permitiera asociar a través de una herramienta "*BackOffice"* el precio pagado por acción con la cotización real y así establecer un seguimiento a largo plazo con el objetivo de determinar de forma

automática el resultado final de la operación sin tener que establecer un seguimiento manual.

También sería interesante explorar las operaciones de venta. Como se ha podido demostrar, los *"insiders"* son fundamentalmente vendedores de acciones y por ese motivo sería conveniente desarrollar un modelo predictivo usando "*machine learning*" con el objetivo de determinar en qué momento estas operaciones de venta generan un hecho significativo que pueda ser útil para un inversor.

6 BIBLIOGRAFÍA

Another competitor says it may sue to stop FirstEnergy's PPA. (03 de 12 de 2015). Obtenido de https://www.bizjournals.com/: https://www.bizjournals.com/columbus/blog/ohio-energy-inc/2015/12/another-competitor-says-it-may-sue-to-stop.html

Barabone, A. (30 de 04 de 2020). *What Investors Can Learn From Insider Trading.* Obtenido de Investorpedia: https://www.investopedia.com/articles/02/061202.asp

Bill, K. (04 de 02 de 2020). *Sarbanes Oxley Act*. Obtenido de Investorpedia: https://www.investopedia.com/terms/s/sarbanesoxleyact.asp

First World Pharma. (15 de 06 de 2010). Obtenido de https://www.firstwordpharma.com/node/666723?tsid=17

fundinguniverse. (06 de 11 de 2020). Obtenido de http://www.fundinguniverse.com/company-histories/roanoke-electric-steel-corporation-history/

Gabele, B. (19 de 06 de 1999). *TheStreet.* Obtenido de https://www.thestreet.com/: https://www.thestreet.com/opinion/when-insiders-exercise-options-look-out-756167

García, L. A. (01 de 01 de 2016). *TFC. Gestión de proyectos.* Obtenido de UOC: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/45590/7/lameijideTFC0116memoria.pdf

Imclone stock trading case. (2020). Obtenido de Wikipedia: https://en.wikipedia.org/wiki/ImClone_stock_trading_case

Investment Intelligence from Insider Trading. (2000). Obtenido de Amazon: https://www.amazon.com/Investment-Intelligence-Insider-Trading-Press/dp/0262692341

Moreland, J. (2000). *Profit from Legal Insider Trading.* Chicago: Dearborn.

Motley Fool. (05 de 01 de 2008). Obtenido de The Motley Fool. (2020). « Why Micron Technology, Inc. Shares Rose 88% in 2017: https://www.fool.com/investing/2018/01/05/why-micron-technology-inc-shares-rose-88-in-2017.aspx

Rainardi, V. (2008). *Building a Data Warehouse.* New York: Springer-Verlag New York, Inc.

Roberts, L. (21 de 10 de 2017). *Americans are still terrible at investing, annual study once again shows.* Obtenido de MarketWatch: https://www.marketwatch.com/story/americans-are-still-terrible-at-investing-annual-study-once-again-shows-2017-10-19

"SEC".GOV. (s.f.). Obtenido de "SEC".GOV: https://www."SEC".gov/interps/telephone/cftelinterps_"SEC"16.pdf

"SEC".gov. (09 de 03 de 2004). Obtenido de "SEC".gov: https://"SEC".report/Document/0000084278-04-000002/

The Motley Fool. (17 de 10 de 2020). Obtenido de The Motley Fool: https://www.fool.com/investing/2020/10/17/why-buying-cheap-stocks-is-the-wrong-investing-

str/?source=awin&awc=12195_1605879211_4bd03ee3537e8ab64f0a7e1d204adcde&utm_source=aw&utm_medium=affiliate&utm_campaign=101248

Why Micron Technology, Inc. Fell 60% in 2015. (05 de 01 de 2016). Obtenido de The Motley Fool: https://www.fool.com/investing/general/2016/01/05/why-micron-technology-inc-fell-60-in-2015.aspx

Wikipedia. (16 de 11 de 2020). Obtenido de https://en.wikipedia.org/: https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008

Wikipedia. (16 de 11 de 2020). Obtenido de Wikipedia: https://en.wikipedia.org/wiki/MeadWestvaco

Wikipedia. (16 de 11 de 2020). Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Talen_Energy

Wikipedia. (19 de 11 de 2020). Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Cree_Inc.

Wikipedia. (2020). *Bear Market (2008-2009)*. Obtenido de https://en.wikipedia.org/wiki/United_States_bear_market_of_2007%E2%80%932009

Wikipedia. (2020). *Black Monday (2011).* Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Black_Monday_(2011)

Wikipedia. (07 de 10 de 2020). *"EDGAR" System.* Obtenido de https://en.wikipedia.org/wiki/"EDGAR": https://en.wikipedia.org/wiki/"EDGAR"

Wikipedia. (2020). *Peter Lynch.* Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Peter_Lynch

7 GLOSARIO DE TÉRMINOS

Término	Descripción
Insider	Persona o entidad jurídica que forma parte de la empresa o tiene alguna participación directa o indirecta en la compañía.
Insider Trading	Negociación de productos financieros de una empresa, como por ejemplo, acciones, realizada en posesión de material público o no público de la misma y cuya negociación es realizada por directivos o entidades con intereses en la empresa.
Issuer	Empresa en la que se realiza la actividad de "*Insider Trading*".
Report Owner	Véase Insider.
Transaction	Detalle de la operación realizada durante la actividad de "*Insider Trading*". Fundamentalmente es una operación de venta o de compra.
SEC	Siglas que corresponden a la "*Securities Exchange Comision*". Es el organismo que se encarga proteger a los inversores así como mantener el orden y el funcionamiento de los mercados financieros.
Form 4 filing	Documento presentado en la "SEC" por un "*insider"* o por un representante legal de este que contiene información sobre el estado o cambio de estado en la propiedad de algún activo financiero de la empresa, como por ejemplo, la compra de acciones.
Accession number	Identificador único de un documento "*form 4*" en el sistema "EDGAR"
CIK	"*Central Index Key*". Identificador único para "*Issuers*" y "*Report Owners*".
Crash	Estallido de una burbuja económica o financiera.
Spin-off	Mediante la separación de una subdivisión de una empresa, se crea una nueva.
CEO	Chief Executive Officer (Principal rol de dirección en la gestión de una compañía)

CFO	Chief Financial Officer (Principal rol de dirección en el apartado financiero de una compañía)
ETL	Abreviación de "*Extract, transform and load"* que es la técnica utilizada para la extracción, transformación y carga de información procedente de multitud de fuentes en una base de datos.
RSS	"*Really Simple Syndication*". Se utiliza para la distribución de un determinado contenido a clientes que se han suscrito.
IDX	Servicio del sistema "EDGAR" para la descarga de documentos "*form 4*" a través de un índice. Permite la descarga de documentos de forma diaria, mensual o anual.
Data warehousing	Es el proceso de construcción y uso en un "*data warehouse*".
Data wrangling	Es el proceso de transformar y mapear datos de un formato base a otro formato.
Data cleaning	Es el proceso de detectar y corregir (o eliminar) datos imprecisos o corruptos en los registros de un "*dataset*".
Python	Lenguaje de programación interpretado que soporta la orientación a objetos.
Pandas	Biblioteca de software para el lenguaje Python para la manipulación y el análisis de datos.
XML	Lenguaje de marcas semiestructurado utilizado para almacenar datos de forma legible.
XSD	Lenguaje para describir la estructura y restricciones del contenido de los datos contenidos en un archivo "XML".
JSON	Formato de texto para el intercambio de datos.
CSV	Formato de texto abierto para la representación de datos en formato de tabla, separando las columnas con un separador determinado.
Dataset	Colección de datos.
Jupyter Notebook	Entorno informático para la ejecución de código en Python. Usado principalmente para el análisis de datos.

8 ANEXO

8.1 ENTREGABLES DEL PROYECTO

- Documentos *"form 4"* (2003-2020): Enlace de descarga
- Backup de la base de datos MariaDB: Enlace de descarga
- Software "ETL" desarrollado en Python: Enlace de descarga
- Datos CSV para la sesión de exploración inicial de datos: Enlace de descarga
- Fichero de sesión de exploración inicial de datos en Jupyter: Enlace de descarga
- Documento de instalación y configuración del proyecto: Enlace de descarga