دوفصلنامه محاسبات و سامانه‌های توزیع شده
www.jdcs.ir

روش کشف تقلب در پایگاه‌داده گرافی

الهام عبدالهی[1]، فرشته آزادی پرند*[1]، علی اصغر صفائی[2]
گروه رایانه، دانشکده علوم ریاضی رایانه، دانشگاه علامه طباطبایی[1]
گروه انفورماتیک پزشکی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران[2]

چکیده

امروزه با توجه به گسترش ارتباطات بایستی حجم عظیمی از اطلاعات تولید، مبادله و در نهایت ذخیره‌سازی شود. سیستم‌های پایگاه‌داده‌ی رابطه‌ای، قابلیت ذخیره‌سازی این حجم از اطلاعات را ندارند. مفهوم NoSQL برای مواجهه با این مشکل در سال ۱۹۹۸ توسط کارلو استروزی ظهور پیدا کرد. از انواع پایگاه داده‌های NoSQL می‌توان به پایگاه داده‌های سندگرا، کلید مقدار، ستون محور و در نهایت پایگاه‌های گرافی اشاره کرد. در این کار تمرکز ما بر روی پایگاه داده‌های گرافی خواهد بود. پرکاربردترین پایگاه داده‌های گرافی، پایگاه‌داده neo4j است. پایگاه‌داده مذکور توسط شرکت‌های معروفی از جمله cisco , hp ,Microsoft ,adobe و نظایر آن استفاده می‌شود. از مهم‌ترین کاربردهای پایگاه‌داده مذکور می‌توان به مساله کشف تقلب و کلاهبرداری، کار با اینترنت اشیا و سیستم‌های توصیه‌گر، اشاره کرد. در این کار به بررسی تقلب افزایش قیمت سهم در یکی از شرکت‌های وابسته به سازمان بورس پرداخته خواهد شد. روند کار به این شکل است که عوامل دخیل در افزایش قیمت سهم شناسایی شده وداده‌های مربوطه در پایگاه‌داده‌گرافی neo4j ذخیره شده است. در نهایت زبان پرس وجوی cypher جهت بازیابی زنجیره‌های تقلب به کار گرفته شده است.

کلمات کلیدی: پایگاه داده غیررابطه‌ای گرافی Neo4j ، زبان پرس وجوی Cypher، کشف تقلب، NoSQL

Email: parand@atu.ac.ir

تاریخچه مقاله:

تاریخ ارسال:۹۷/۱۱/۱
تاریخ اصلاحات:۹۸/۳/۱
تاریخ پذیرش: ۹۸/۴/۱
تاریخ انتشار: ۹۸/۵/۱۵

Keywords:

Graph database
Fraud detection
NoSQL database
Cypher query language

Fraud Detection Method in NoSQL Graph Database

Elham Abdollahi[1], Fereshteh-Azadi Parand[*1], Ali Asghar Safaei[2]
[1]Allameh Tabataba'i University, Tehran
[2]Department of Medical Informatics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran

Abstract

Today, with the expansion of communications, a huge amount of production, exchange and ultimately data storage must be stored. Relational databases do not have the capacity to store this volume of information. The NoSQL concept emerged in 1998 by Carlo Strouzi to deal with this problem. NoSQL database types can include documentary databases, value keys, column-driven and eventually graphic databases. In this work, our focus will be on the graphic database. The most widely used graphic database is the neo4j database. The database is used by famous companies such as Cisco, HP, Microsoft adobe and the like. One of the most important uses of the database is the issue of fraud and fraud detection, work with the internet of things and recommendation systems. This work will investigate the fraud in the share price increase in one of the affiliated companies of the stock exchange. The market price is fluctuating momentarily. But this price change is not always based on the value of the relevant share. There are a number of factors involved in the story that have increased share prices without any added value. This paper examines the factors involved in increasing the share price and preserving relevant data in the neo4j database. Finally, the cypher query language is used to retrieve cheat chains.

روش ارجاع به مقاله: ا.عبدالهی، ف.آزادی پرند، ا.صفایی، روش کشف تقلب در پایگاه‌داده‌های گرافی دوفصلنامه محاسبات و سامانه های توزیع شده ، سال دوم، شماره اول، سال ۱۳۹۸،ص ۴۴-۵۴

۱ - مقدمه

امروزه اغلب سازمان‌ها درگیر مشکلاتی از قبیل تضاد منافع، اخاذی اقتصادی، رشوه خواری، پولشویی، پرداخت‌های ساختگی و مسائلی ازاین قبیل هستند. این موارد همگی تقلب یا کلاهبرداری محسوب می‌شوند. طبق تعریف انجمن بازرسان رسمی تقلب[۱]، تصرف افراد در منافع سازمان جهت غنی‌سازی خود، تقلب است که زیان‌هایی با هزینه‌های گزاف را برای سازمان متحمل می‌کند. آنچه باعث تقلب می‌شود، بصورت اضلاع مثلث تقلب با عناوین فرصت، فشار و منطق‌تراشی دسته‌بندی شده است. اما برای بازیابی چنین تقلب‌هایی باید به بررسی اطلاعات ذخیره شده در پایگاه‌داده سازمان پرداخته شود. پایگاه‌داده‌ها در دونوع رابطه‌ای و غیر رابطه‌ای[۱] دسته بندی می‌شوند.

امروزه با گسترش ارتباطات، ذخیره‌سازی و بازیابی‌داده‌ها واطلاعات در پایگاه‌داده‌های رابطه‌ای، بدلیل تاثیر عواملی چون حجم زیاد، تنوع، سرعت تغییر بالا، دقت اطلاعات و یک سری عوامل دیگر، امری هزینه بر است. مشکل پایگاه‌داده‌های رابطه‌ای این است که هنگام پیوست جداول در پرس‌وجوها مجموعه بزرگی ازاطلاعات را بدون فیلترکردن نمایش می‌دهند. درنتیجه سرعت کم و درمقابل پیچیدگی زمانی بالایی دارند. داده‌ها با گذشت زمان تغییر می‌کنند. تغییرات داخلی و خارجی یک سیستم و زمینه‌ای که در آن کار می‌شود، می‌تواند تاثیر قابل‌توجهی را در سرعت داشته باشد. عامل دیگر که باعث عدم عملکرد بهینه و مناسب پایگاه‌داده‌های رابطه‌ای شده است، تنوع در ساختارداده ها است. تنوع، به درجه نظم یا بی‌نظمی، تراکم یا عدم تراکم، اتصال یا عدم اتصال داده‌ها مربوط می‌شود. با این وجود نکته این است که اطلاعات در طبیعت لزوما به صورت جدولی نیست. فرض کنید که در یک رابطه دوستانه در یک شبکه اجتماعی بخواهیم دوستان دوستان دوستان را بازیابی کنیم. چنین پرس وجویی در پایگاه داده‌ی رابطه‌ای حجم عظیمی کد ایجاد کرده(پیچیدگی محاسباتی وفضایی دارد) و انسجام را از بین می‌برد. لذا زمان انتقال به یک پایگاه‌داده‌ی غیررابطه‌ای فرارسیده است. از انواع پایگاه‌داده‌های غیررابطه‌ای می‌توان به پایگاه‌داده‌های سندمحور، کلید مقدار، ستون‌گرا وگرافی اشاره کرد که هریک معماری و ویژگی عملیاتی متنوعی دارند و متناسب با منطق کسب وکار مورد استفاده قرار می‌گیرند.

درپایگاه داده‌های سندگرا اسناد به صورت جفت کلید_مقدار ذخیره می‌شوند. این مقادیر همانند فرمت‌هایی از جمله JSON و XML می‌توانند به صورت لیست یا نقشه باشند. اسناد می‌توانند با شناسه، ذخیره و بازیابی شوند. به عبارت بهتر یک پایگاه‌داده سندگرا، می‌تواند مشابه پایگاه‌داده کلیدمقدار عمل کند. در موارد کلی، پایگاه داده‌های سند محور به شاخص‌ها می‌پردازند تا دسترسی به اسناد را براساس هریک از ویژگی‌های آنها سرعت بخشند. در این صورت با نوشتن شاخص داده، می‌توان به جزئیات داده‌ها دسترسی داشت.

درپایگاه‌داده‌های کلید_مقدار، مقادیر مبهم ذخیره شده، با کلید بازیابی می‌شوند. هرداده‌ای با درهم‌سازی شناسه، (کلید) ذخیره می‌شود. با فرض این که چند باکت و چند ماشین داشته باشیم، تابع درهم‌ساز طوری ساخته می‌شود که توزیع یکنواخت درباکت در دسترس را فراهم کند. به طوری که هیچ ماشین واحد به یک نقطه کانونی تبدیل نشود. با توجه به کلید، می‌توان از

این آدرس برای ذخیره مقدار در باکت مربوط استفاده کرد.

پایگاه داده‌ی ستون‌گرا، پایگاه‌داده‌ای است که مدل داده ای آن بر پایه جدولی با جمعیت کم است. یک ردیف می‌تواند شامل هر تعداد ستون دلخواه باشد. ساده ترین واحد ذخیره سازی ستون است که شامل یک جفت نام و مقدار می‌باشد. ترکیب هر تعداد ستون را ابرستون گویند. هرابرستون یک نام مختص به خود را دارد. اگر چندین ستون کنار هم در یک ردیف ذخیره شوند، به آن خانواده ستون گویند. اگرچندین ابرستون دریک ردیف قرارگیرند، به آن خانواده ابر ستون گویند.

پایگاه داده‌های گراف به عنوان معروف‌ترین ابزار ذخیره سازی و پرس‌وجو در گراف‌ها هستند. پایگاه‌داده‌های گراف برای حل اطلاعات پیچیده و مشکلات تجاری حائز اهمیت هستند. گراف‌ها یک مدل طبیعی برای طیف وسیعی از حوزه‌ها هستند. به عنوان نمونه برای توصیف نحوه عبور کاربر از نقطه A به نقطه B پایگاه داده‌گرافی راه حل مناسبی است.[۲] انتخاب یک پایگاه‌داده غیررابطه‌ای وجایگزینی احتمالی آن با یک پایگاه‌داده رابطه‌ای، به دلیل عدم همسویی مناسب با نیازهای موجود یکی از اتفاقات مهمی است که این روزها در بسیاری از سازمان‌ها در حال وقوع است.

تقلب در سازمان‌های بیمه و بانک باعث از دست‌رفتن سالیانه میلیاردها دلار می‌شود. صاحبان کسب وکار بایستی راه‌حل‌ها و یا سیستم‌های مناسبی را جهت کشف تقلب در نظر بگیرند. امروزه با توجه به گسترش اطلاعات و ارتباطات، روابط بین سازمان‌ها از طریق شبکه‌های اجتماعی صورت می‌گیرد. برای ذخیره‌سازی اطلاعاتی از نوع شبکه‌های‌اجتماعی نیاز به پایگاه‌داده های گرافی است. مناسب‌ترین پایگاه‌داده‌گرافی جهت ذخیره‌سازی شبکه‌های اجتماعی، پایگاه داده neo4j است. یکی از کاربردهای پایگاه‌داده مذکور، کشف تقلب سازمان‌ها است[۳]. تقلب‌های تجاری در مواردی رخ می دهد که کارمند، مدیر، افسر و یا مالک سازمان به نفع خود و ضرر آن سازمان تقلب کند. سه نوع عمده تقلب تجاری عبارتند از : فساد، مصادره دارایی و بیانیه‌های دروغین. فساد می‌تواند در تضاد منافع که ریشه در طرح خرید وفروش دارد، رشوه خواری، پاداش غیرقانونی و اخاذی اقتصادی دیده‌شود. منظور از دارایی، نقدینگی و موجودی انبار و دارایی‌های دیگراست. مصادره دارایی به معنای سرقت از نقدینگی و پرداخت‌های ساختگی به روش‌های مختلف است. درصورتی که نسبت سود به درآمد بیش از حد ویا حتی بسیارکمتر بیان شود، این بیانیه‌ای دروغین است که کلاس دیگری از تقلب است. مدیران می‌بایست نتایج عملکرد خود را به مالکان واحد تجاری و سایر تأمین‌کنندگان وجوه مانند بانک‌ها گزارش کنند. بااین حال، گزارش‌گیری مالی شامل طیف وسیعی از اعداد حسابداری برای نشان دادن سود، جریان نقدی و وضعیت مالی واحد تجاری است. بنابراین، مدیریت می‌تواند با استفاده نادرست از حسابداری ساختگی، وضعیت مالی شرکت را بهتر از واقعیت نشان داده و از اثرات آن در کوتاه مدت بهره ببرد[۴]. در این کار به بررسی تقلب‌های از نوع بیانیه دروغین پرداخته خواهد شد. تقلب مذکور توسط هریک از عوامل مدیر، کارمند ومشتری قابل انجام است. این نوع تقلب‌ها که در اثر مخفی کردن واقعیات و یا حتی ارائه اطلاعات گمراه کننده در گزارش‌ها طی دوره‌های مختلف یک شرکت است، باعث ایجاد یک سری منافع

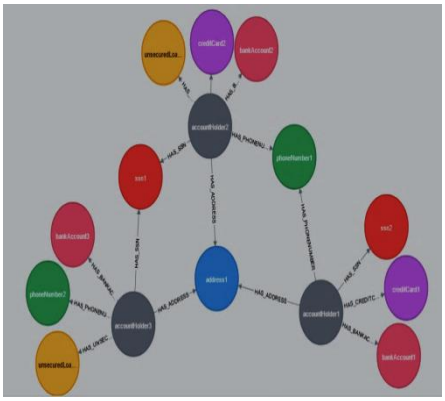
مالی برای متقلبان و زیان‌هایی سنگین برای سرمایه گذاران دیگر می‌شود.

از طریق طراحی پرس‌وجوهای مناسب و بررسی دقیق روابط بین موجودیت‌ها، می‌توان الگوی مناسبی را جهت کشف تقلب در نظر گرفت تا آنچه مطابق با این الگو است، از درون پایگاه‌داده‌گرافی بازیابی کرد. در ادامه کارهای مرتبط انجام شده، الگوریتم کشف تقلب در افزایش قیمت سهم یک شرکت بورسی بصورت فلوچارت بیان شده است. در نهایت نتایج کار را خواهیم داشت.

۲ – کارهای مرتبط انجام شده

یکی از رویکردهای کشف تقلب که در پایگاه داده های غیررابطه‌ای مانند neo4j بررسی می‌شود، استفاده از مدل گرافی است. بدین شکل که با پیمایش گره‌ها و روابط بین آنها و نیز خواص گره های مذکور، می‌توان الگوهایی را برای شناسایی حلقه‌های تقلب درنظر گرفت. براساس این الگوها، می‌توان پرس وجوهای مناسبی را طراحی کرده و در نهایت حلقه تقلب مذکور را بازیابی کرد.

مدل کشف تقلب در سیستم بانکی در[۴] الگویی که مربوط به کشف تقلب سیستم بانکی است، مطرح شده است. مدلسازی این مساله بصورتی است که دارندگان حساب بانکی که بصورت غیرمستقیم و از طریق دسترسی به اطلاعات مخفی دیگر دارندگان حساب، با یکدیگر در ارتباط هستند، به عنوان گره‌های متقلب شناسایی و بازیابی می شوند. ساختار کلی گراف عملکرد سیستم بانکی مطابق با شکل ۱ است. گره‌های طوسی رنگ دارندگان حسابی هستند که به اطلاعات افراد دسترسی دارند.



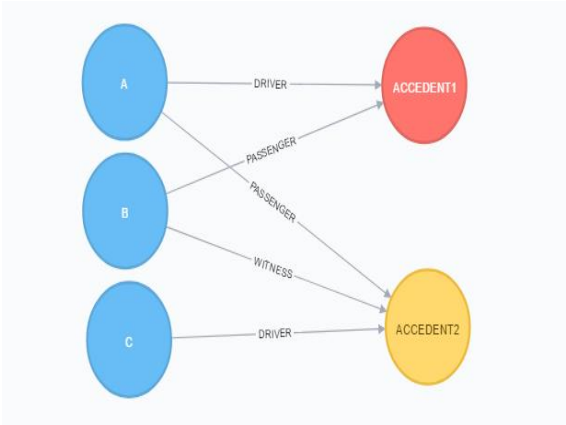
شکل ۱ .گراف دسترسی افراد به اطلاعات در سیستم بانکی

در ادامه گام‌هایی برای پرس‌وجوی مناسب جهت یافتن حلقه‌های تقلب، بررسی شده است. در ابتدا دارندگان حسابی که به اطلاعات دسترسی دارند، جست وجو می شوند. تعداد دارندگان حساب تحت عنوان اندازه حلقه تقلب مد نظر قرار می‌گیرد. دارندگان حسابی که از طریق داشتنcreditCard و یا بدست آوردن unsecuredLoan ، حق دسترسی به حساب‌های دیگر را کسب کرده‌اند، جست وجو شده و هریک با شناسه منحصر بفرد خود گردآوری می‌شوند. درصورتی که دسترسی از طریق ارتباط HAS_CREDITCARD صورت گرفته باشد، مجموع حدپایین حساب و درصورتی که دسترسی از طریق ارتباط HAS_UNSECUREDLOAN صورت گرفته باشد، مجموع حد تعادل حساب به عنوان ریسک مالی در نظر گرفته می‌شود.

مدلی برای کشف تقلب سازمان بیمه

یک مدل تقلب در سازمان بیمه حالتی است که متقلبان با صحنه‌سازی حوادث جعلی به سازمان مربوطه مراجعه کرده و ادعای خسارت می‌کنند. مدل

تقلب به حالتی است که افرادی که در یک حادثه هستند، درحادثه دیگر نیز حضور داشته باشند. با این تفاوت که نقش افراد متفاوت باشد. به عنوان نمونه فردی با برچسبA در حادثه ۱ به عنوان DRIVER و در حادثه۲ به عنوان PASSENGER باشد. فرد B در حادثه ۱ نقش PASSENGER و در حادثه ۲ تحت عنوان WITNESS عمل کرده است. لازم به ذکر است در حادثه ۲ فرد دیگری(C) تحت عنوان DRIVER وارد عمل شده است. شکل ۲ نحوه عملکرد را نمایش می‌دهد.



شکل۲ . نمونه ای از تقلب در سازمان بیمه

پرس وجوی مربوطه جهت یافتن حلقه‌های تقلب به این شکل صورت می‌گیرد که گره‌های با برچسبA,B,C که به هر دو گره با برچسبACCEDENT1,2 متصل شده‌اند، به عنوان گره‌های مقلب شناسایی وبازیابی می‌شوند.[۵]

مدل کشف تقلب تجارت الکترونیک

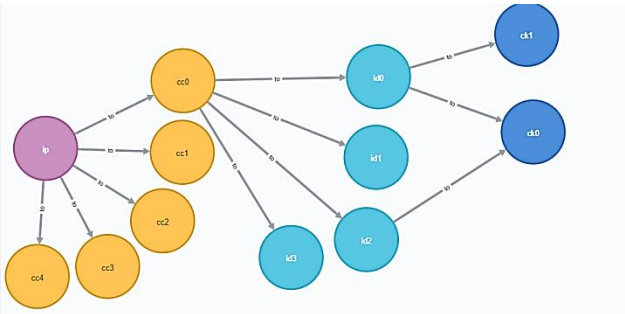
همانطور که زندگی ما بطور فزاینده‌ای به سمت دیجیتالی شدن پیش می‌رود، تعداد معاملات مالی برخط[۶]، رشد می‌یابد. متقلبین بسرعت خود را با این روند وفق داده و روش‌های زیرکانه‌ای برای فریب سیستم پرداخت برخط، طراحی کرده‌اند. این فعالیت می‌تواند حلقه‌های جنایی را درگیر کند. متقلبین ماهر می‌توانند تعداد بسیار زیادی از هویت‌های ساختگی ایجاد کرده و از این حالت برای پیاده‌سازی طرح‌های بسیار بزرگ استفاده کنند.

معاملات برخط با شناسه‌هایی که در ادامه بررسی شده است، در نظرگرفته می‌شود: شناسه کاربر، آدرس ip، محل جغرافیایی، شماره کارت اعتباری و ردیابی کوکی‌ها. به طور معمول انتظار می‌رود روابط بین این شناسه‌ها بطور نسبی به حالت یک به یکی نزدیک باشد. بعضی از تغییرات به طور طبیعی برای حساب‌های مشترک، دستگاه‌ها و یا استفاده مشترک اعضای خانواده‌ای از یک کارت اعتباری و یا استفاده افراد از چندین کامپیوتر و مانند این موارد قابل تحمل هستند. با این حال، به محض اینکه روابط به سمت عددی فراتر از حالت معقول رفت، اغلب تقلب در حال وقوع است. ممکن است برای نمونه، تعداد زیادی از کاربران معاملاتی ناشی از ip یکسان داشته باشند. درتعداد بسیاری از معاملات، برای آدرس‌های مختلف، از یک کارت اعتباری استفاده شده باشد. تعداد زیادی از کارت‌های اعتباری همگی برای یک آدرس استفاده شده باشد. در هریک از سناریوها، الگوهای درون گراف، با گذراز روابط بین قطعات مجزا، از طریق اطلاعاتی که می‌توانند تحت عنوان نشانه‌های قوی رویداد تقلب به کار گرفته شوند، قابل بازیابی هستند. عامل بزرگی که به عنوان دغدغه محسوب می‌شود، روابط بیش‌تر در میان شناسه‌ها است. گراف‌های بهم پیوسته شاخص‌های قدرتمندی برای انجام چنین تقلب‌هایی هستند.

پایگاه داده‌های گراف برای الگویابی در زمان واقعی طراحی شده‌اند. با انجام بررسی‌ها در محل و ارتباط آنها با محرک‌های رویداد، چنین طرح‌هایی قبل

از اینکه بتوانند خسارت قابل توجهی ایجاد کنند، می توانند کشف شوند. محرک، می‌تواند رویدادهایی مانند ورود به سیستم، ثبت سفارش یا ثبت نام کارت اعتباری را شامل شود.

گراف شکل ۳، تراکنش آدرس ip با رویداد تقلبی که از ip رخ می‌دهد را نشان می‌دهد.cc_x نشانگر شماره کارت اعتباری، id_x نشانگر idکاربری برای انجام تراکنش‌های بر خط وck_x به کوکی مشخص ذخیره شده در سیستم اشاره می‌کند. در شکل ۳، ip مربوطه، با استفاده از ۵ کارت اعتباری، چندین معامله را انجام می‌دهد که یکی از آنها(cc_0)، توسط چندین id استفاده شده است. دو کوکی ck_1,ck_0 بین دو id مشترک است.

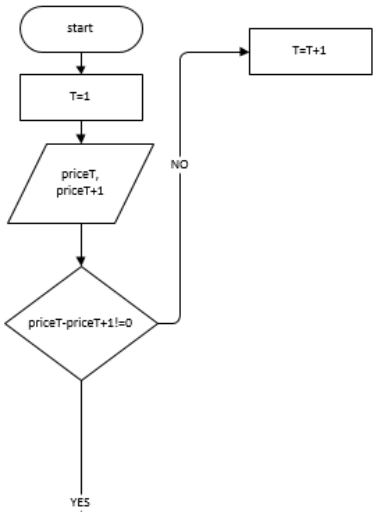


شکل ۳ . نمونه‌ای از تقلب تجارت الکترونیک

۳- روش پیشنهادی

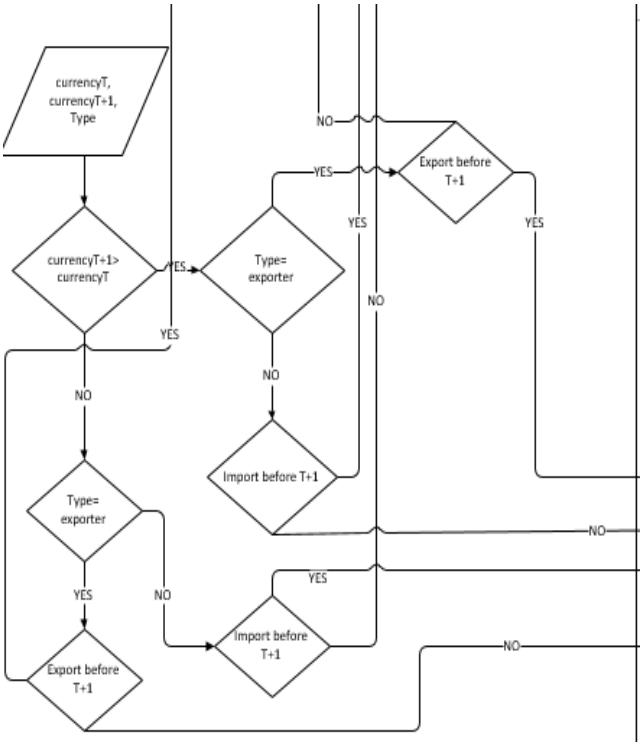
- الگوریتم کشف تقلب قیمت سهم

زمانی که قیمت سهم تغییر پیدا می‌کند، بایستی یک سری عوامل بررسی شوند. ازجمله این عوامل می‌توان به وضعیت ارز، تصمیمات دولت، روابط بین‌الملل، میزان عرضه وتقاضا، وابستگی صنعت به صنعت دیگر، میزان انحصاری بودن، نرخ تورم و میزان دارایی شرکت مورد نظر اشاره کرد. تک تک این عوامل را بصورت گره یک گراف در نظر می‌گیریم. طبق شکل ۴، ابتدا قیمت سهم در دو لحظه متوالی دریافت می‌شود. در صورت افزایش قیمت سهم وارد مرحله بررسی عامل ارز می‌شویم. در غیر اینصورت قیمت سهم در دولحظه متوالی دیگر دریافت می‌شود.



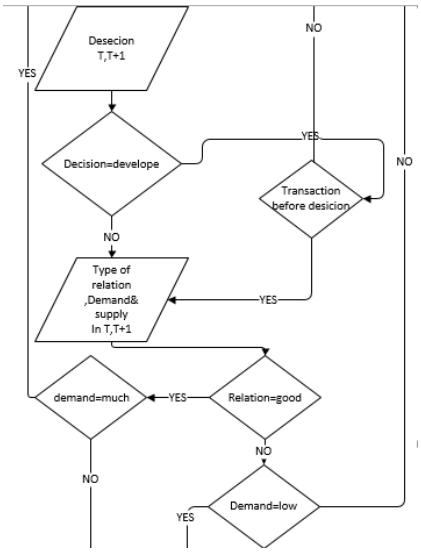
شکل۴. بخشی از فلوچارت بررسی اختلاف قیمت سهم در دولحظه متوالی

بایستی نوع شرکت از نظر وارد کننده یا صادر کننده بودن مشخص شود. در صورت افزایش قیمت ارز و صادرکننده بودن، افزایش قیمت سهم طبیعی است. در صورت افزایش قیمت ارز و وارد کننده بودن، افزایش قیمت سهم عامل دیگری دارد. پس باید عامل تصمیمات دولت بررسی شود. در صورت کاهش قیمت ارز و صادرکننده بودن، افزایش قیمت سهم عامل دیگری دارد و باید تصمیمات دولت بررسی شود. در صورت کاهش قیمت ارز و وارد کننده بودن، افزایش قیمت سهم طبیعی است. بخشی از فلوچارت مربوطه در شکل ۵ آورده شده است.



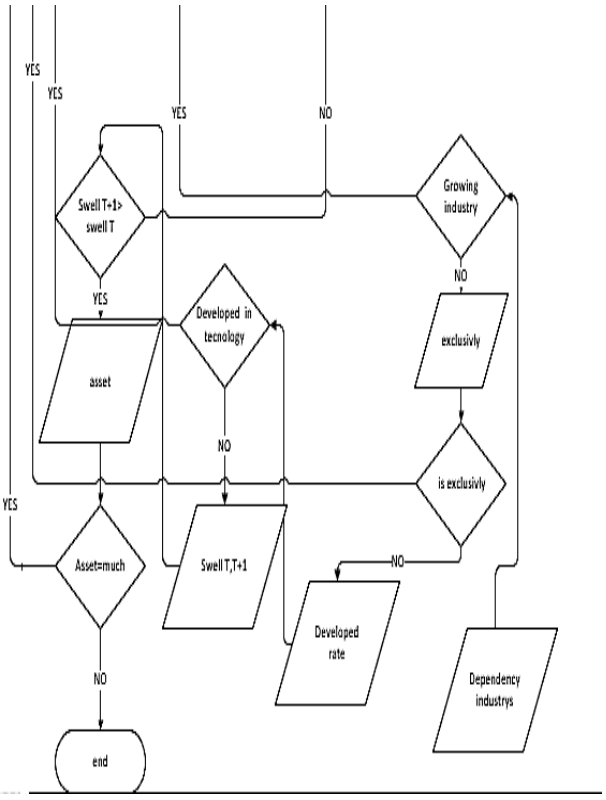
شکل ۵. بخشی از فلوچارت بررسی وضعیت ارز

اگر دولت تصمیمی مبنی بر افزایش قیمت سهم گرفته باشد، افزایش قیمت سهم طبیعی است و قیمت دو لحظه متوالی دیگر بررسی می‌شود. در غیر اینصورت عامل روابط بین الملل و میزان عرضه وتقاضا بررسی می شود. بخشی از فلوچارت مربوطه در شکل ۶ آورده شده است.



شکل ۶. بخشی از فلوچارت روابط بین الملل

نحوه روابط به هرشکلی باشد باید میزان تقاضا بررسی شود. در صورت زیاد بودن تقاضا، افزایش قیمت سهم طبیعی است. پس قیمت در دو لحظه متوالی دیگر بررسی می‌شود. در غیر اینصورت وضعیت صنایع وابسته بررسی می‌شود.



شکل۷ . بخشی از فلوچارت مربوط به سایر عوامل تاثیر گذار بر افزایش قیمت سهم

وضعیت صنایع وابسته بررسی می‌شود. درصورت روبه رشد بودن صنایع وابسته، افزایش قیمت سهم طبیعی است و قیمت دو لحظه متوالی دیگر بررسی می‌شود. در غیر اینصورت عامل انحصاری بودن شرکت، بررسی می شود. در صورت انحصاری بودن شرکت، افزایش قیمت سهم طبیعی است. در غیر اینصورت عامل نرخ تورم بررسی می‌شود. در صورت افزایش نرخ تورم، افزایش قیمت سهم طبیعی است. در غیر اینصورت عامل دارایی شرکت بررسی می‌شود. در صورتی که دارایی شرکت

زیاد باشد، افزایش قیمت سهم طبیعی است در غیر اینصورت تقلب رخ داده است. بخشی از فلوچارت مربوطه در شکل ۷ آورده شده است.

پرس وجوی مربوطه جهت بازیابی حلقه تقلب در ادامه بررسی شده است. برای هر گره به ترتیب زمان یک id در نظر می‌گیریم. پرس وجوی لازم برای بازیابی گره آغازگر تقلب در ادامه بررسی می‌شود. روند کار به این شکل است که گره‌های متوالی بررسی می‌شوند. پس باید گره‌هایی که اختلاف id آنها یک واحد است با هم مقایسه شوند. حال به بررسی خصوصیات دو گره متوالی پرداخته خواهد شد. خاصیت قیمت گرهی با id بزرگتر باید بیش‌تر باشد، چراکه در این کار به بررسی تقلب افزایش قیمت سهم پرداخته شده است. وضعیت ارز بررسی می‌شود. قیمت ارز گرهی با id کمتر باید بیش‌تر ازقیمت ارز گره دیگر باشد. خاصیت عرضه وتقاضا بررسی می‌شود. میزان عرضه وتقاضا گرهی با id کمتر به ترتیب باید کم‌تر و بیش‌تر از میزان عرضه وتقاضا گره با id بیش‌تر باشد. خاصیت تورم بررسی می‌شود. نرخ تورم گرهی با id کمتر باید بیش‌تر ازنرخ تورم گرهی با id بیش‌تر باشد. در نهایت خاصیت دارایی بررسی می‌شود. میزان دارایی گرهی با id کمترباید بیشتر از میزان دارایی گرهی با id بیش‌تر باشد. اگر تمامی خصوصیات گرهی با id کم‌تر برابر با خصوصیات گره باid بیش‌تر باشد، دلیلی برای افزایش قیمت سهم وجود ندارد. گره مربوطه باید به عنوان تقلب بازیابی شود. برای نمایش بهتر گره متقلب در پایگاه داده گراف، گره مربوطه را با یک ارتباطی به نامfraud به خودش متصل کرده و تمامی گره‌هایی که ارتباط fraud داشته باشند، به عنوان گره آغازگر تقلب بازیابی می‌شوند.

۴ – نتایج

داده‌های این کار شامل اطلاعات مالی یک شرکت بورسی صادرکننده است. از جمله داده‌های مالی به کارگرفته شده می توان به قیمت سهم در زمان‌های مختلف، قیمت ارز، تغییرات نرخ تورم، زمان صادرکردن، میزان عرضه و تقاضا برای کالاهای شرکت، وضعیت صنایع وابسته به کالاهای شرکت، میزان انحصاری بودن، تصمیمات دولت، میزان پیشرفت وتکنولوژی و مجموع دارایی‌های شرکت که برای سال‌های ۱۳۹۴و۱۳۹۵، در نظر گرفته شده است، اشاره کرد. البته به دلیل عدم تاثیر، یک سری از عوامل مذکور در این نوع شرکت، نادیده گرفته شده است. از آن جمله می‌توان به تصمیمات دولت، وضعیت صنایع وابسته(که هر سال بدتر شده است، لذا نمی‌تواند تاثیری در افزایش قیمت سهم داشته باشد) و میزان انحصاری بودن اشاره کرد. چراکه شرکت مربوطه انحصاری نبوده وفضا رقابتی است و افزایش قیمت سهم با درجه بسیار کم‌تری رخ می‌دهد. البته امکان دسترسی به یک سری از اطلاعات به دلیل محرمانه بودن وجود نداشت. (از جمله زمان صادرات)

این روش برای محاسبه خطایی که در افزایش قیمت سهم رخ می‌دهد، استفاده می‌شود. شاید به نظر برسد که چه خطایی در کاهش قیمت سهم ممکن است باشد، اما باید دانست یک سری از شرکت‌ها برای فرار از مالیات قیمت‌های خود را پایین می‌آورند تا خود را زیان ده جلوه دهند. ابتدا داده‌های مربوطه موجود در فایل اکسل به فرمت csv تبدیل شده و در پایگاه داده neo4jبارگزاری می‌شود. هر رکورد فایل بصورت یک گره ذخیره می‌شود. هریک از داده‌های مربوطه بصورت

خاصیت رکورد در نظر گرفته می‌شود. برای مشاهده میزان ثابت بودن قیمت‌ها، گره‌های با قیمت یکسان را با یک ارتباط به هم متصل می‌کنیم. که پرس‌وجوی مربوط به این کار به ترتیب در شکل های ۸و۹و۱۰نمایش داده شده است.

```
match(n),(m)
where
tointeger( n.id)+1= tointeger(m.id)
and
tointeger(n.price)=tointeger(m.price)
merge (n)-[:to]->(m)
```

شکل۸ .پرس و جوی مربوطه جهت بارگزاری داده

```
USING PERIODIC COMMIT 1000
load csv with headers from "file:///c:/test.csv" as line
create
(:share{id:line.id,date:line.date,time:line.time,
price:line.price,type:line.type,
currency:line.currency,swell:line.swell,
ir:line.ir,supply:line.supply,
demand:line.demand,darayee:line.darayee})
```

شکل ۹ .پرس و جوی مربوط جهت ایجاد روابط

```
match(n),(m)
where
(tointeger( n.id)+1= tointeger(m.id) )
and (tointeger(n.price)<tointeger(m.price))
and(tointeger(n.currency)>tointeger(m.currency))
and (tointeger(n.supply)<tointeger(m.supply))
and (tointeger(n.demand)>tointeger(m.demand))
and (tointeger(n.swell)>tointeger(m.swell))
and (tointeger(n.darayee)>tointeger(m.darayee))
merge(m)-[r:fraud]-(m)
return r as fraud
```

شکل ۱۰. پرس و جوی مربوط جهت بازیابی گره متقلب

زمان لازم برای بارگذاری ۲۴۸۱۶ رکورد و تنظیم ۲۴۸۱۵۶ خاصیت تنها ۴ ثانیه محاسبه شده است. در ضمن برای ادغام یک سری گره‌ها برای ایجاد ۲۴۳۱۸ رابطه ۲۳ ثانیه زمان محاسبه شده است. طبق آزمایش‌های انجام شده، ۴۹۸ مورد تغییر قیمت داشته ایم که شامل هردوحالت افزایش و کاهش در قیمت‌ها می‌باشد. البته در این کار تقلب‌های مربوط به افزایش قیمت سهم با میزان دقت ۰,۹۷ بازیابی می‌شود. تمامی ۱۴۳ مورد تقلب واقعی همه کشف شده‌اند. برای پیاده سازی کار از سیستمی با پردازنده intel(r) core(TM) i7-5500u cpu@2.40GHz 2.40GHz و RAM 6.00GB استفاده شده است و پایگاه داده neo4j مورد استفاده نسخه ۳,۵,۰ آن است.

۵- مراجع

[1] Han, Jing, et al. "Survey on NoSQL database." 2011 6th international conference on pervasive computing and applications. IEEE, 2011.
[2] L., Webber, J. & Eifrem, E. (2015). Graph databases: new opportunities for connected data. " O'Reilly Media, Inc." 1-64.
[3] www.neo4j.com
[4] Rathle, S., Rathle, G.& Rathle, Ph. (2014). Fraud detection: Discovering connections with graph databases." White Paper-Neo Technology-Graphs are Everywhere,1-11.
[5] Laurent, C., Laurent, A. & Laurent, A. (2016) "Rogue behavior detection in NoSQL graph databases." Journal of Innovation in Digital Ecosystems 3(2), 70-82.
[6] Rathle, S., Rathle, G.& Rathle, Ph. (2014). Fraud detection: Discovering connections with graph databases." White Paper-Neo Technology-Graphs are Everywhere,1-1

الهام عبداللهی عابد مدرک کارشناسی خود را در رشته علوم کامپیوتر در سال ۱۳۹۴ از دانشگاه دولتی محقق اردبیلی و مدرک کارشناسی ارشد خود را در همان رشته گرایش سیستم‌های کامپیوتری در سال۱۳۹۸ از دانشگاه دولتی علامه طباطبائی اخذ کرده است.

فرشته آزادی پرند کارشناسی خود را از دانشگاه علم و صنعت، کارشناسی ارشد را از دانشگاه تربیت مدرس و دکتری خود را از دانشگاه علم و صنعت در رشته مهندسی کامپیوتر نرم افزار اخذ نموده است.
ایشان از سال ۱۳۹۱ به عنوان هیئت علمی گروه رایانه در دانشگاه علامه طباطبایی مشغول به کار هستند. زمینه های پژوهشی مورد علاقه ایشان سیستم ها و پایگاه داده‌های توزیعی و تئوری بازی می باشد.

علی اصغر صفائی متولد سال ۱۳۵۷ کارشناسی و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر گرایش نرم افزار به ترتیب در سال‌های ۱۳۸۰ و۱۳۸۳ به پایان رسانید. او در سال ۱۳۹۰ موفق به دریافت درجه دکترا در همین رشته و در حوزه پایگاه داده‌ها از دانشگاه علم وصنعت ایران شد. وی هم‌اکنون به‌عنوان عضو هیئت علمی گروه انفورماتیک پزشکی دانشگاه تربیت مدرس مشغول به فعالیت است. زمینه‌های پژوهشی ایشان شامل سامانه‌های پایگاه‌داده و جریان داده، داده‌های عظیم، پردازش موازی و بی‌درنگ پرس‌وجوها، حافظه‌های نهان معنایی، امنیت در سامانه‌های پایگاه‌داده و مدیریت اطلاعات در مقیاس وب است. از ایشان کتاب و مقالات متعددی نیز در همین زمینه‌ها به چاپ رسیده است.