فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Summer 2021, 2 (2), 43-54
DOR: 20.1001.1.27832570.1400.2.1.6.7

Received: 21 Mar. 2021
Accepted: 25 May 2021

Use of ICD (International Classification of Diseases) to prepare medical data in decision support systems

Mahmood Moradi[1], Mehdi Afzali[2*], Forough Sadat Hosseini[3]

1. Assistant Professor, Department of Information Science and Dentistry, Razi University, Kermanshah, Iran. mahmoudmoradi@razi.ac.ir
2.Assistant Professor, Department of Information Technology Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran. *(Corresponding Author)* afzali@iauz.ac.ir
3. Zanjan University of Medical Sciences, Zanjan, Iran.

Abstract

Introduction: Increasing business intelligence informs the organization about the information of the business environment and provides the possibility of correct and timely analysis of data and information. Business intelligence is an umbrella term that was presented by Howard Dresner in 1989 to define a set of concepts and methods to improve business decision-making using a reality-based support system. The intelligent business system is an application that is employed to better understand the organization's data in the process of making detailed decisions. The structure of business intelligence includes five layers. The process of extracting data from various sources, and converting and loading them into the data warehouse is generally called ETL. These different sources can include XML data sets, relational tables, non-relational sources, weblogs, spreadsheets, etc. In the ETL layer, transformation is the most important process because at this stage, data errors are corrected so that it can be used in managers' decisions after loading it into the data warehouse. This research was done with the aim of describing and explaining the data cleaning method to remove bad data.

Method: The proposed method is designed with the help of one of the visual programming languages called C-Sharp version 2013 and SQL Server 2014, and in order to test the performance of the program, it is used from the databases of people who died due to illness. Different methods were used in 2013. To prepare this program, C# and SQL commands were used.

Finding: After the process of error detection and correction of information banks, the modified information bank is evaluated by the designed program to check the amount of correction of contaminated data in ILTEC architecture and ETL. Apllying on 12,424 records, it showed had an error of 0.008%, which was not related to the implementation method and was the user's error at the time of data entry that a national code was registered for two different people. But in the ETL architecture, compared to the ILT architecture, there are still data that are not found in the troubleshooting stage to be corrected.

Conclusion: In order to test the IT architecture, from a database with 5464 records (diseases of Zanjan province), 10203 records (diseases of two provinces of Zanjan and Ilam) and 14739 records (diseases of three provinces of Zanjan, Ilam and Hamedan) was used (it was used in the ILTEC architecture). It took 102 seconds to run the process of finding errors and correcting the program with the database, which had 5464 records, and 1.78% of the errors were in the database. In order to test again, using the data bank that had 10203 records, this time the time of executing the error detection and correction process was 154 seconds, and the number of errors was 2.65%. Finally, we used the database that had 14,739 records, the execution time of which was 293 seconds, and the number of errors in the process of error detection and correction was 2.3%.

Keywords: Business Intelligence (BI), Data Preparation, Data warehouse, ICD Code, ILTEC.

فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

استفاده از طبقه‌بندی بین‌المللی آماری بیماری‌ها(ICD) برای آماده‌سازی داده‌های پزشکی در سیستم‌های پشتیبان تصمیم

تاریخ دریافت: ۱۳۹۹/۱۲/۰۲
تاریخ پذیرش: ۱۴۰۰/۰۲/۲۵

سال دوم، تابستان ۱۴۰۰
شمارهٔ دوم، صص: ۴۳-۵۴

محمود مرادی[۱]، مهدی افضلی[۲*]، ،فروغ السادات حسینی[۳]

۱. استادیار، گروه علم اطلاعات و دانش شناختی، دانشگاه رازی، کرمانشاه، ایران. mahmoudmoradi@razi.ac.ir
۲. استادیار، گروه مهندسی فن‌آوری اطلاعات، واحد زنجان، دانشگاه آزاد اسلامی، زنجان، ایران. (نویسنده مسئول) afzali@iauz.ac.ir
۳. دانشگاه علوم پزشکی و خدمات درمانی و بهداشتی، زنجان، ایران.

چکیده: کیفیت اطلاعات در موفقیت تجزیه و تحلیل اطلاعات بسیار حیاتی و مهم است. اطلاعات بارگذاری‌شده در انبارداده باید صحیح، دقیق و با کیفیت باشد. داده با کیفیت در انباره داده موجب تحلیل مناسب و تصمیم‌گیری بهتر می‌شود. همچنین مباحث کیفیت داده باید قبل از بارگذاری در انباره داده مورد توجه قرارگیرد. پاکسازی داده به مفهوم یافتن و حذف خطاها است. همچنین در این فرایند داده‌های اضافی و ناسازگار شناسایی می‌شوند. پاکسازی‌داده در مرحله استخراج، انتقال، بارگذاری با اطمینان از کیفیت داده در انباره‌داده موجبات اثربخشی هوشمندی کسب‌وکار را فراهم‌می‌آورد. هدف پاکسازی‌داده، شناسایی داده‌های بد (اشتباه، نامرتبط و ناقص) به منظور اصلاح یا حذف آن‌ها است تا از دقت و سازگاری مجموعه داده اطمینان حاصل شود. این پژوهش با هدف تشریح و تبیین روش پاکسازی داده برای حذف داده‌های بد انجام شده‌است. بانک‌اطلاعاتی نمونه از اطلاعات بیماری‌های استان‌های زنجان، ایلام و همدان تشکیل‌شده است. به منظور حل مشکلات داده در بانک نمونه از فرم سی‌شارپ و ابزاهای نرم‌افزار اس‌کیوال استفاده‌شده‌است. بخش اصلی نتایج نشان‌می‌دهد که به‌کارگیری روش پاکسازی داده موجب کاهش میزان خطای بانک داده تا میزان ۰/۰۰۸ درصد شده‌است.

واژه‌های کلیدی: هوش‌تجاری، آماده‌سازی داده، انباره داده، ICD، ILTEC.

۱. مقدمه

در سال‌های اخیر، هوشمندی کسب‌وکار به یکی از مفاهیم اساسی مدیریت تبدیل‌شده و در سازمان‌های پیشرو، با فرهنگ سازمانی عجین شده‌است. افزایش هوشمندی کسب‌وکار، سازمان را نسبت به اطلاعات محیط کسب‌وکار آگاه نموده، امکان تجزیه و تحلیل صحیح و به‌موقع داده‌ها و اطلاعات را فراهم‌می‌آورد. نتایج حاصله در قالب فرم‌ها و گزارش-های مناسب ذخیره و در مواقع مقتضی به عنوان حق انتخاب برای تصمیم‌سازی بیشتر، در دسترس مدیران قرارمی‌گیرد. از این طریق جریان تبادل اطلاعات و دانش در بستر سازمان تسریع شده، کارایی و اثربخشی در خصوص فرآیند تفکر جمعی و تصمیم‌گیری بهبود می‌یابد. بیشترین بهره‌مندی به‌دست‌آمده از هوشمندی کسب‌وکار، امکان دسترسی بی‌واسطه، به داده‌ها توسط تصمیم‌گیرندگان در تمام سطوح سازمان است. در این صورت این افراد قادرند که با داده‌ها تعامل داشته-باشند و آن‌ها را تحلیل‌کنند و در نتیجه کسب‌وکار را مدیریت‌کنند، کارایی را بهبود بخشند، فرصت‌ها را کشف کنند و کارشان را با بازدهی بالا انجام‌دهند.

۲. هوش تجاری

با افزایش مقدار اطلاعات تولیدشده توسط سازمان فرآیند تصمیم‌گیری بسیار پیچیده‌تر می‌شود و نیاز به زمان زیادی خواهدداشت. بنابراین بهتر است در اجرای تصمیم‌گیری از تکنولوژی هوشمندی کسب‌وکار استفاده-شود[1].

هوشمندی کسب‌وکار چتری است که توسط هوارد[۱] درسنر در سال ۱۹۸۹ برای تعریف مجموعه‌ای از مفاهیم و روش‌هایی جهت بهبود تصمیم‌گیری کسب‌وکار با استفاده از سیستم پشتیبانی مبتنی بر واقعیت ارائه‌شد. گوشال[۲] و کیم[۳] اولین تعریف علمی را در مورد هوش ارائه‌نمودند. طبق تعریف این دو، هوش به‌عنوان یک فلسفه مدیریت و ابزاری جهت کمک به مدیریت سازمان می‌باشد و اطلاعات کسب‌وکار را به منظور ایجاد تصمیم‌های مؤثر اصلاح‌می‌نماید [2]. طبق نظر اکرسون[۴] هوشمندی کسب‌وکار باید بتواند ابزارهای مثل گزارش تولید، پرس‌وجوی کاربر نهایی، فرآیند پردازش آنلاین[۵]، ابزار داشبور/ صفحه نمایش، ابزار داده‌کاوی[۶] و ابزار مدل‌سازی و برنامه‌ریزی فراهم‌نماید [3, 4].

سیستم هوشمند کسب‌وکار نرم‌افزاری کاربردی است که برای فهم بهتر داده‌های سازمان در فرآیند تصمیم‌گیری‌های دقیق استفاده‌می‌شود. ساختار هوشمندی کسب‌وکار شامل پنج لایه می‌باشد:

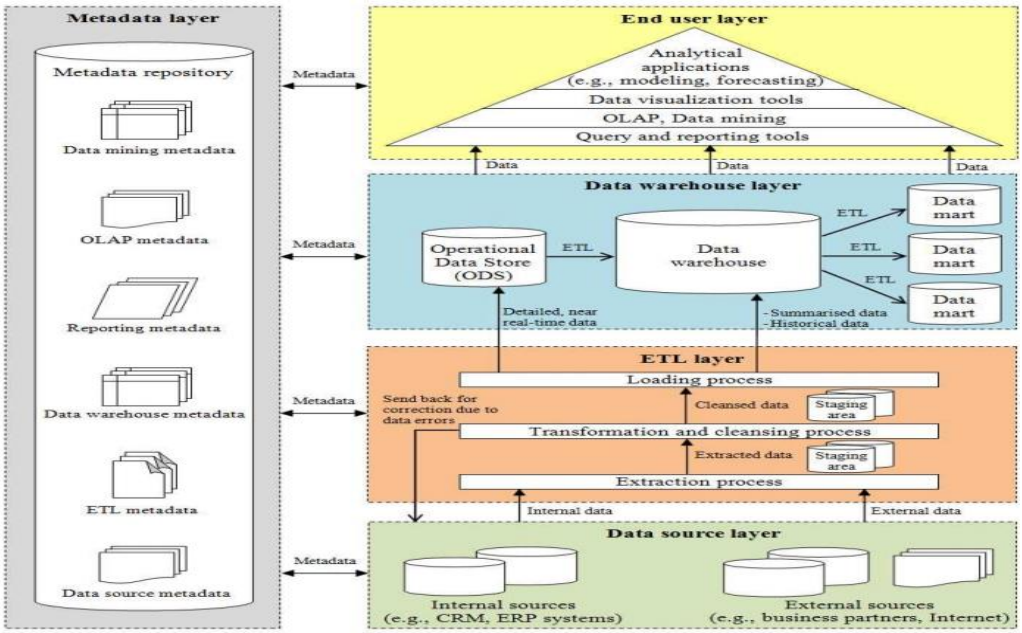
۱-لایه منبع داده[۷]

۲- لایه ای تی ال [۸](استخراج[۹]، تبدیل[۱۰]، بارگذاری[۱۱])

۳- لایه انباره‌داده[۱۲]

۴- لایه فراداده[۱۳]

۵- لایه کاربرنهایی[۱۴] [5, 6, 7] (نگاه شود به شکل ۱)



شکل ۱.: معماری سیستم هوش تجاری[7]

۱،۲. لایه منبع داده

لایه منبع داده شامل داده‌های ساختاریافته، غیرساختاریافته و نیمه ساختار یافته است. همه این داده‌ها می‌توانند از دو منبع داخلی و خارجی فراهم‌شوند. منابع اطلاعات داخلی به سیستم‌های عملیاتی داخلی سازمان اشاره می‌کند و منابع داده خارجی به منابعی که از خارج سازمان نشأت می‌گیرد، اشاره‌دارد[5, 6, 7].

۲،۲. لایه ای تی ال (استخراج، تبدیل، بارگذاری)

فرآیند استخراج داده از منابع مختلف، تبدیل و بارگذاری آن‌ها به داخل انباره‌داده عموماً، ای تی ال نامید می‌شود. که این منابع مختلف می‌تواند

شامل مجموعه داده‌های ایکس‌ام‌ال، جداول رابطه‌ای، منابع غیر رابطه‌ای، وب لاگ، صفحات گسترده و غیره باشد[4].

"ای تی ال" فرآیندی است که مسئول خارج‌کردن داده‌های خام از سیستم منابع مختلف، تبدیل به فرمت مشترک و سپس قراردادن داده‌ها در انباره داده‌ها است[5, 7, 8]. این لایه داده‌ها را از سیستم‌هایی که به‌صورت فرآیند تراکنشی برخط می‌باشند به انباره‌داده انتقال‌می‌دهد. همچنین می‌تواند برای انتقال اطلاعات از یک انباره‌داده به انباره‌داده دیگر نیز استفاده‌شود[7, 8]. بررسی داده، تشخیص، اجرا، کنترل بعد از پیش پردازش،[11] جریان بازگشتی داده‌های پاکسازی‌شده به ترتیب مراحلی است که در فرآیند اصلاح اجرامی‌شود[9].

از مشکلات مرحله پاکسازی می‌توان به ناسازگاری مقادیری که در فیلدهای دارای معنی یکسان ذخیره‌شده‌اند[5] (تضاد ساختاری[4])، تکرار داده‌ها، داده‌های اشتباه، [5] وجود مقادیر غیرقابل‌قبول،[4, 5] تضاد نام‌گذاری، بررسی مقادیر از دست رفته[4] اشاره‌نمود.

داده از منابع گوناگون در فرمت‌های مختلف جمع‌آوری‌می‌شود بنابراین فرمت استاندارد و پاکسازی داده‌ها از جمله نیاز محیط انباره‌داده می‌باشد. در واقع، این فرآیند مقادیر، نوع فیلدها، ساختار و غیره را بررسی‌می‌کند و بعد از مشخص‌شدن داده‌های معیوب و تکراری[10]، به‌وسیله ابزارهای برنامه‌نویسی سنتی، زبان‌های اسکریپت‌نویسی یا زبان پرس‌وجوی ساختار یافته آن‌ها را اصلاح‌می‌کند[5, 7] .

داده‌های حاصل از این فرآیند به منظور کشف فرآیند دانش باید خصوصیاتی از قبیل دقت، صحت،[4, 9, 11] به‌موقع بودن،[4] یکپارچه و کامل‌بودن،[9, 11] مطابقت نمایش، [9] تراکم، یکتایی، هم‌ریختی [9, 11] و سازگاری را داشته‌باشند[11].

۳،۲. لایه انباره‌داده

مخزن مرکزی برای ذخیره داده‌های تولیدشده در لایه ای‌تی‌ال است. امروزه از آنجاکه انباره‌داده به سطح بالایی از بازدهی رسیده‌است، فرصت‌های جدیدی در استخراج اطلاعات از بانک اطلاعاتی قابل‌استفاده، وجود دارد[12]. انباره داده یک سازمان بزرگی از پایگاه داده‌ها است، که وظیفه اصلی آن انتشار اطلاعات سازمان به طور مؤثر به منظور پشتیبانی استراتژیک از تصمیم‌گیری مدیران سازمان است. درواقع این بخش مسئول نگهداری همه کپی‌هایی از داده‌های منابع مختلف سازمان است.-[13]. بیل اینمون و رالف کیمبال دو پیشگامان در نظریه ساخت انباره داده‌ها می‌باشد[14]. طبق نظر رالف کیمبال انباره‌داده یک کپی از سیستم‌های تراکنشی به‌ویژه ساختاریافته برای فرآیند پرس‌وجو و تجزیه و تحلیل است[15]. از ویژگی‌های اصلی انباره‌داده می‌توان به موضوع-گرا[۱۵]، یکپارچه، متغیر با زمان[۱۶]، از بین نرفتنی[۱۷] اشاره‌کرد[5, 6, 16].

جداول در انباره‌داده می‌تواند به شکل‌های مختلفی طراحی شوند یکی از این مدل‌ها، مدل بعدی[۱۸] است که برای غلبه به عملکرد پرس‌وجو در داده‌های بزرگ در انباره‌داده استفاده‌می‌شود. این مدل، عملکرد پرس-وجو را برای تهیه گزارش‌های خلاصه بهبود بخشد و به فضای بیشتری نسبت به مدل رابطه‌ای نیاز دارد. این مدل دارای جدول واقعی[۱۹] و ابعادی[۲۰] می‌باشد.

جداول واقعی شامل کلیدی برای ارتباط با جداول ابعادی می‌باشد. جداول ابعادی شامل جزئیاتی درباره جداول واقعی است. این جداول شامل اطلاعات توصیفی در مورد مقادیر عددی جداول واقعی است.[18, 17, 16] ویژگی‌هایی که برای یک سازمان در طراحی مهم است، جزء جداول بعدی قرارمی‌گیرد. درواقع اساس مدل ابعادی است[19] و می-تواند به‌صورت ستاره‌ای[۲۱]و یا دانه برف[۲۲] طراحی‌شود[16, 17].

در این لایه سه جزء انباره‌داده، داده عملیاتی و انبارک قراردارد. جریان داده از انباره‌داده عملیاتی به انباره‌داده و سپس به انبارک‌ها می-باشد[5]. با توجه به این که داده‌های موجود در انباره‌داده اساساً برای پشتیبانی از نیازها در کل سازمان استفاده می‌شود، بنابراین برای پشتیبانی از نیازها و درخواست‌های یک واحد سازمانی خاص تجهیز نشده‌اند. انبارک زیر مجموعه انباره‌داده است و همانند انباره شامل داده-های تاریخی است که به کاربران اجازه‌می‌دهد که به روند داده‌های مختلف دسترسی داشته باشند و تحلیل‌کنند [5]. اما حجم داده ذخیره شده در آن خیلی کمتر از داده ذخیره‌شده در انباره داده است [5, 20].

۴،۲. لایه فراداده

فراداده راجع به داده است. به این معنی که به داده‌هایی که برای توصیف داده‌ها به‌کار می‌روند فراداده گفته‌می‌شود. فراداده توصیف می‌کند که داده‌ها کجا استفاده و ذخیره می‌شوند، منبع داده کجا است، چه تغییراتی در داده صورت گرفته‌است و یک جزء داده چگونه با بقیه اطلاعات مرتبط است. مخزن فراداده محلی برای ذخیره‌سازی داده‌های مربوط به فراداده است[5, 6, 13].

انواع مختلف فراداده همانند منابع داده، ای‌تی‌ال، گزارش‌گیری، اولپ، داده‌کاوی وجوددارد [5, 13].

۵،۲. لایه کاربر نهایی:

لایه کاربرنهایی شامل ابزارهایی برای نمایش اطلاعات به شکل‌های مختلف به کاربران متفاوت است[5, 6, 7] از ابزارهای آن می‌توان ابزارهای پرس‌وجو وگزارش‌گیری، اولپ، تصویرسازی[۲۳]، ابزارهای تحلیلی، مدیریت عملکرد کسب‌وکار[۲۴] را نام برد[5].

۳. اهمیت موضوع

در لایه ای تی ال، تبدیل مهمترین فرآیند است زیرا در این مرحله خطاهای داده اصلاح می‌شود تا در نهایت بعد از بارگذاری در انباره‌داده، در تصمیم‌های مدیران مورد استفاده قرارگیرد.

بانک اطلاعاتی می‌تواند فیلدهایی از قبیل: نام، نام‌خانوادگی، جنسیت، تاریخ تولد، سن، آدرس، کدپستی، کدملی، شهر، استان، نام بیماری، علائم بیماری، داروی تجویزشده، دوز دارو و غیره را داشته‌باشد که در هنگام ادغام بانک‌های اطلاعاتی و یا هنگام ثبت توسط کاربر دچار خطا شود. همچنین ممکن است در دو بانک اطلاعاتی مختلف فیلد مشترکی وجود داشته‌باشد ولی عنوان فیلد و نحوه ورود آن‌ها متفاوت

باشد که بعد از ادغام باید به یک مدل در بانک اطلاعاتی ویرایش‌شود. (مثل جنسیت که در بانکی زن/مرد ثبت‌شده و در بانک دیگری ۱/۰)

به‌علت خطای کاربر و یا ادغام بانک‌های اطلاعاتی، ممکن است بانک جدید حاوی رکورد تکراری و یا فیلدهای ثبت‌نشده در محیط استقرار موقتی پایگاه داده به‌وجودآید. لذا نیاز است که خطاهای موجود، با سرعت بیشتری اصلاح‌گردد و سپس به انباره‌داده بارگذاری‌شده تا در تصمیم-گیری مدیران به‌طور صحیح استفاده‌شود.

سؤال‌های پژوهش عبارتنداز:

- چگونه می‌توان مرحله ای تی ال را با سرعت بالاتری اجرا نمود؟
- چگونه می‌توان داده‌های آلوده بانک‌های اطلاعاتی را جهت فرآیند تصمیم‌گیری اصلاح نمود؟
- چگونه می‌توان مقادیر ثبت نشده بانک‌های اطلاعاتی را جهت فرآیند تصمیم‌گیری پرنمود؟

۴. پیشینه تحقیق

دارشان تانک و همکارانش در سال ۲۰۱۰ بیان‌می‌کنند که دو عملگر پیوند[۲۵] و تجمیع[۲۶] که نقش اساس در پیش‌پردازش جهت دستکاری و تحکیم داده‌ها برروی انباره‌داده بازی‌کرده و استفاده از آن‌ها سبب صرفه-جویی در زمان می‌شود[8]. السایاق و همکارانش در سال ۲۰۱۱ دیاگرام نگاشت موجودیت (ای‌ام‌دی[۲۷]) را به عنوان مدل ادراکی[۲۸]جدید برای فرآیند ای‌تی‌ال ارائه‌نموده‌اند[21].

حسین طالب‌زاده در سال ۲۰۱۲ ضمن بررسی مشکلات ابزارهای ای‌تی‌ال موجود به معرفی سرویس‌های الگوی ای تی ال برمبنای فراداده می‌پردازد وی خلاصه‌ای از انواع مختلفی از فراداده و کاربردهایشان را بیان‌کرده‌است[13].

پروال و همکارانش در سال ۲۰۱۳ به شرح عملکرد دو الگوریتم هدکلین و پیوستگی در پاکسازی داده‌ها پرداختند و در انتها این دو الگوریتم را باهم مقایسه نمودند[22]. بهاتاچارچی و همکارانش در سال ۲۰۱۳ فایل متنی که حاوی حروف، اعداد، جنسیت و شناسه می‌باشد را براساس فرآیند ای تی ال اصلاح نمودند[23]. بهاتاچارچی و همکارانش در سال ۲۰۱۳ با ارائه یک سری از الگوریتم‌هایی، روش قبلی خود را بهبودبخشیدند[24]. خدری و همکارانش در سال ۲۰۱۳ به کمک اطلاعات جبری یک نظریه عمومی برای تعریف ساختار مجموعه داده‌ها و پاکسازی ارائه نموده‌اند و در نهایت الگوریتمی برای تولید قوانین انجمنی بر مجموعه داده‌های ساخت‌یافته ارائه کرده‌اند[25].

شوران در سال ۲۰۱۴ تحقیقاتی از سال ۱۹۹۲ تا ۲۰۱۰ بر روی مسائل کیفیت داده در انباره‌داده انجام داده‌است نتایج به‌دست‌آمده از این تحقیقات شامل: داده‌ها هرگز به‌صورت کامل از منابع قبلی خود به‌دست نمی‌آید، سیاست و برنامه‌ریزی مدیریتی وجود ندارد و یکپارچه-سازی سیستم‌ها ناهمگون می‌باشد[26].

اشوینی ساو و همکارانش در سال ۲۰۱۴ برای تمیزکردن داده‌ها روش ترکیبی که شامل الگوریتم بهبودیافته پی‌ان‌آراس (تشخیص و تصحیح داده‌های متنی)، تکنیک پیشرفته (تصحیح داده‌های کمی) و الگوریتم متعدی (حذف رکوردهای تکراری و فیلدهای خالی) است را ارائه‌نموده‌است[27]. پرنا کالکارانی و همکارانش در سال ۲۰۱۴ الگوریتم ترکیبی به نام هدکلین که شامل نسخه‌های اصلاح شده الگوریتم پی‌ان-آراس و الگوریتم بستن متعدی است را ارائه‌نموده‌است[28]. بهاتاچارچی و همکارانش در سال ۲۰۱۴ علاوه بر الگوریتم موجود تمیزکردن داده‌ها مانند پی‌ان‌آراس، الگوریتم‌های مختلف مانند: پاکسازی رشته و عدد، نسبت آمار، بررسی فرهنگ لغت داده، بررسی فراداده و غیره را برروی سیستم پایگاه داده شهروند اجرا نمودند[29].

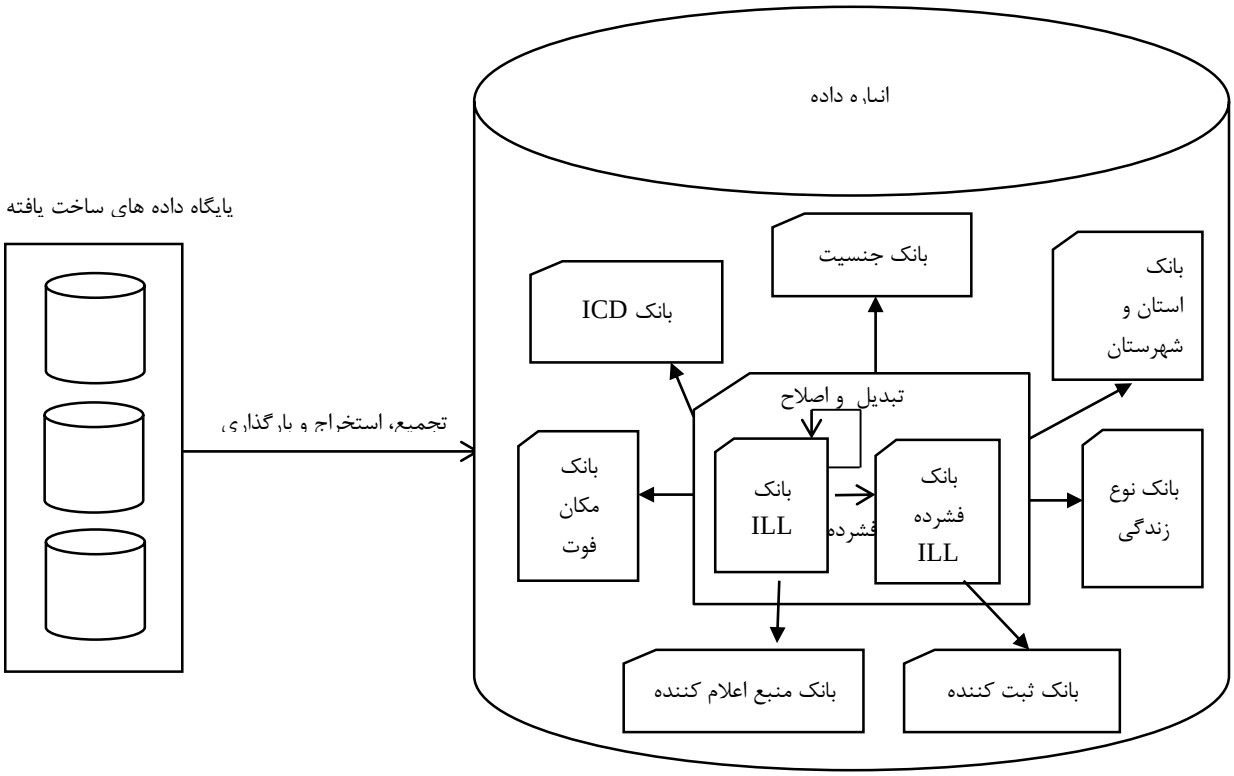
۵. روش پیشنهادی

روش پیشنهادی، به کمک یکی از زبان‌های برنامه‌نویسی ویژوالی بنام سی‌شارپ ورژن ۲۰۱۳ و اس‌کیو ال سرور ۲۰۱۴ طراحی شده‌است و جهت تست عملکرد برنامه، از بانک‌های اطلاعاتی افراد فوت‌شده به علت بیماری‌های مختلف در سال ۱۳۹۳ استفاده‌شد. جهت تهیه این برنامه از دستورهای سی‌شارپ و اس‌کیوال استفاده‌شد. که دستورهای اس‌کیوال سرور در قالب استورپراسیجر[۲۹] مورد استفاده قرارگرفت.

۱،۵. معماری ارائه شده

در معماری ارائه شده به نام «آی‌ال‌تی‌ای‌سی» (شکل ۲)، همه بانک‌های اطلاعاتی ساخت‌یافته را دریافت‌نموده، ادغام‌می‌کند. در زمان ادغام فیلدی به نام "نام استان ثبت‌کننده" و "نام شهرستان ثبت‌کننده" به بانک اطلاعاتی ill اضافه‌نموده و با ادغام هر بانک، نام استان و نام شهرستان ثبت‌کننده را در آن فیلد ایجادشده کپی‌می‌کند.

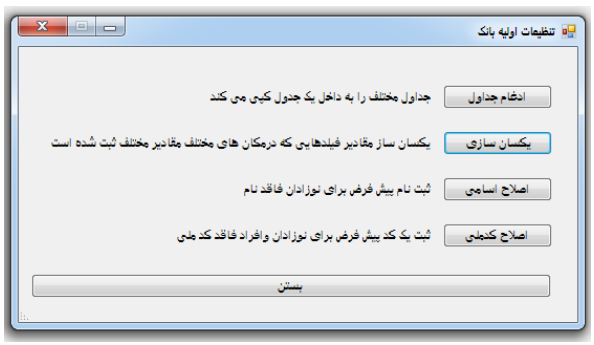
سپس فرآیند یکسان‌سازی را بدون بررسی صحت مقادیر، انجام می-دهد. از آنجاکه نوزادان مرده به دنیا آمده فاقد نام هستند در مرحله بعدی با توجه به جنسیت آن‌ها عبارت" نوزاد دختر" و یا "نوزاد پسر" در فیلد نام ثبت می‌شود. در انتها، کدشناسه را برای فیلد کدملی نوزادان و افرادی که کدملی آن‌ها ثبت شده کپی‌می‌کند.



شکل ۲: معماری پیشنهادی(آی‌ال‌تی‌ای‌سی30)

در این معماری، داده‌ها وارد انباره‌داده شده و سپس فرآیند اصلاح در همان انباره‌داده بدون ناحیه عملیاتی انجام‌می‌گیرد. با این روش مرحله انتقال اطلاعات به محیط عملیاتی در هر لایه حذف‌شده که نسبت به معماری ای‌تی‌ال، که دارای محیط عملیاتی در هر لایه است، سبب افزایش سرعت در اجرای برنامه شده‌است.

مدل بانک اطلاعاتی در داخل انباره‌داده به صورت ستاره می‌باشد یعنی دارای یک جدول واقعی و هشت جدول ابعادی است. که خود این جدول ابعادی از دو قسمت تشکیل شده‌است یک قسمت مربوط به داده-های اصلاح شده با فرمت اولیه و قسمت دیگر شامل همان داده‌ها ولی با فرمت فشرده می‌باشد. بدین ترتیب مدیران می‌توانند به‌سرعت، فرآیندهای آنالیزی خود را انجام‌دهند. در شکل ۳ تنظیمات اولیه‌ای که برای بانک‌های اطلاعاتی اعمال‌می‌شود را مشاهده‌می‌کنید.



شکل ۳. تنظیمات اولیه در برنامه طراحی شده

۲،۵. نحوه عملکرد برنامه:

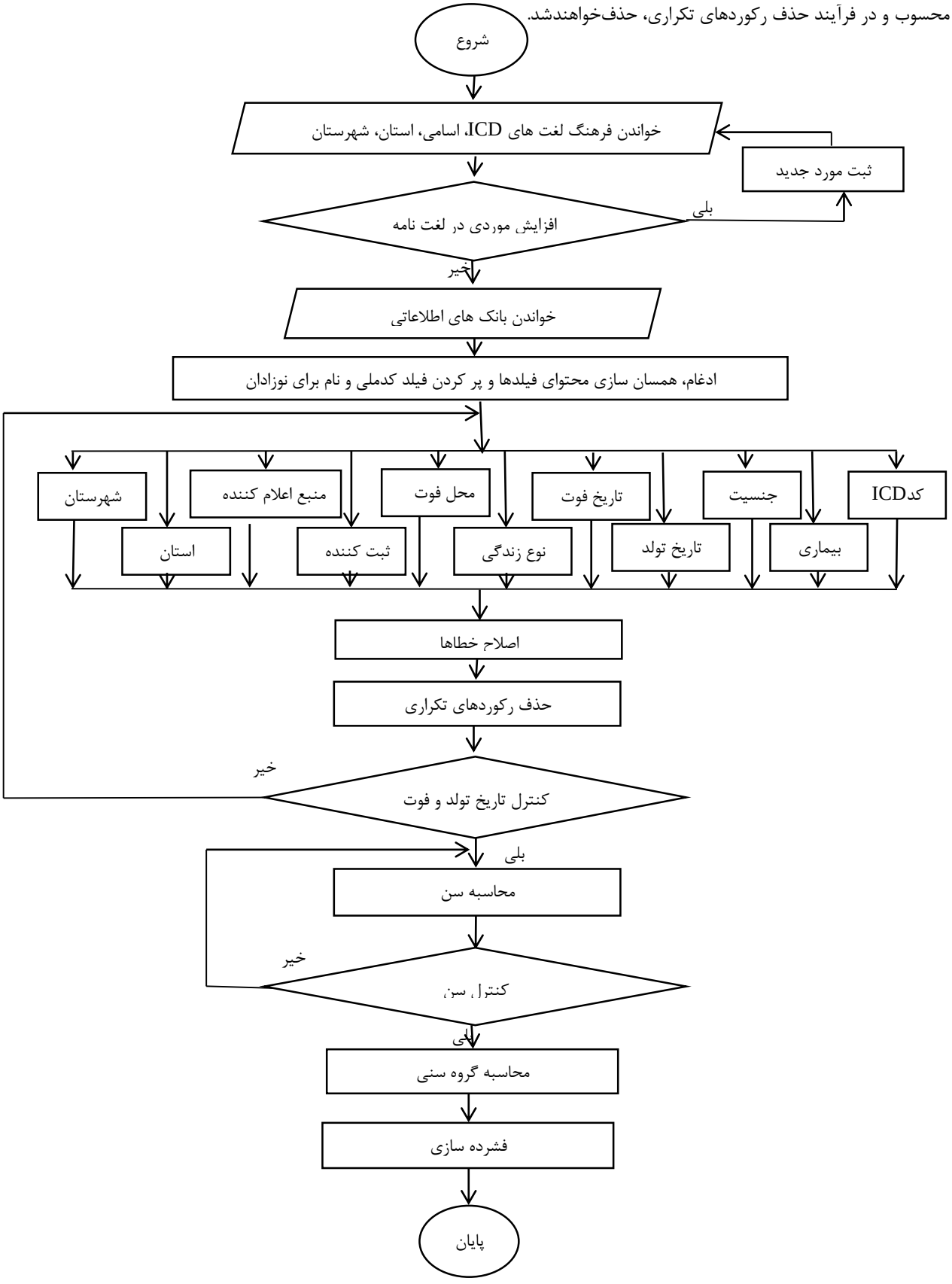
براساس فلوچارت ارائه شده در شکل ۴ ابتدا فایل اکسل افراد فوت‌شده به داخل پایگاه‌داده ایجادشده در اس‌کیوال جداگانه منتقل‌می‌شوند. سپس محتوای بانک اطلاعاتی باهم ادغام شده و در داخل بانک اطلاعاتی ill قرارمی‌گیرد.

در فرآیند ادغام فیلدهای به نام " نام استان ثبت‌کننده " و " نام شهرستان ثبت‌کننده" در بانک اطلاعاتی ایجادمی‌گردد و نام استان و نام‌شهرستان ثبتکننده برای تمام رکورد های آن کپی‌می‌شود.

ازآنجایی که محتوای هر بانک توسط افراد مختلف در مکان‌های مختلف جمع‌آوری‌شده، ممکن است محتوای یک فیلد در یک بانک به-صورت کد و در بانک دیگری به‌صورت رشته ثبت‌شود پس در مرحله بعد، عملیات یکسان‌سازی محتوای فیلدها (از قبیل: جنسیت، نوع زندگی، محل فوت، منبع اعلام کننده، فرد ثبت‌کننده) را براساس محتوای فراداده اجرامی‌کند.

بعد از اجرای فرآیند یکسان‌سازی، حال نوبت به پرکردن فیلد نام و کدملی است. زیرا برای بررسی فیلد جنسیت لازم است که فیلد نام خالی و یا حاوی عبارت بی محتوایی نباشد لذا در این مرحله عبارت "نوزاد دختر" و یا "نوزاد پسر" برای نوزادانی که مرده به‌دنیاآمده و فاقد نام می‌باشند، ثبت‌می‌گردد. کدشناسه (در بانک ill یونیک تعریف شده) در قسمت کدملی برای نوزادان مرده به دنیا آمده و افرادی که به هر دلیل کدملی آن‌ها در فایل اولیه ثبت نشده، ثبت خواهدشد. زیرا بدون انجام

این فرایند رکوردهایی که فاقد کد ملی می‌باشند جزء رکوردهای تکراری محسوب و در فرآیند حذف رکوردهای تکراری، حذف‌خواهندشد.



شکل ۴: فلوچارت عملکرد برنامه

۱،۲،۵. خطایابی در بانک اطلاعاتی

مرحله بعد یافتن خطاهای موجود در بانک ill می‌باشد. این خطاها عبارتند از: جنسیت، تاریخ تولد، نوع زندگی، تاریخ فوت، بیماری، کد ICD، محل فوت، فرد ثبت‌کننده، منبع اعلام‌کننده، استان و شهرستان (محل تولد، زندگی و فوت)، رکورد تکراری و فیلدهای فاقد مقدار.

نوع زندگی: در وزارت بهداشت نوع زندگی بستگی به نوع خدماتی که ارائه‌می‌شود تقسیم‌بندی‌می‌گردد. (شهری و روستایی، که خود روستایی هم شامل روستای اصلی (دارای مرکز بهداشتی درمانی)، روستای قمر(فاقد خانه بهداشت و تحت پوشش خانه بهداشت دیگر)، روستای سیاری (دارای خانه بهداشت و تحت پوشش مرکز بهداشتی درمانی)، روستای غیرساکن). لذا در صورت عدم ثبت و یا داده اشتباه در این فیلد، رکوردهای این گونه فیلدها جزء داده‌آلوده شمرده می‌شوند.

کدICD: هر بیماری در جهان دارای کد بین‌المللی است که به ICD معروف است ممکن است کاربر حین ثبت، کد بیماری را اشتباه ثبت‌کند (بدلیل تعداد زیاد کدها) و یا ثبت‌نکند. لذا دراین مرحله به کمک فیلد نام بیماری و لغت نامه ICD که تهیه‌شده اطلاعات این فیلد بررسی می‌شود در صورت مغایرت جزء داده آلوده در نظر گرفته‌می‌شود.

نحوه محاسبه گروه سنی براساس دستورالعمل وزارت بهداشت به شرح زیر می‌باشد: کمتر از ۳۰ روز، نوزاد ۳۰ روز؛ ۱۱ ماه و ۲۹ روز، شیر خوار یک سال؛ ۵ سال و۱۱ماه و۲۹ روز، کودک ۶ سال؛ ۱۷ ماه و۱۱ماه و ۲۹ روز، نوجوان ۱۸ سال، ۲۹ سال و ۱۱ماه و ۲۹ روز، جوان ۳۰ سال، ۵۹ سال و ۱۱ماه بالای ۶۰ سال، کهن‌سال

در این برنامه از سه لغت‌نامه برای تطبیق محتوای فیلدها (اسامی، جنسیت، استان، شهرستان، بیماری و کدهای ICD) تحت عنوان اسامی، کدهای بین‌المللی ICD و استان/ شهرستان استفاده شده است. بانک اطلاعات اسامی که برای این منظور تهیه شده‌است حاوی۶۷۷۳ نام و بانک اطلاعاتی کدهای بین‌المللیICD ۱۰۸۸۴ بیماری در آن ثبت-شده که در این برنامه امکان افزایش نام، بیماری، استان و شهرستان جدید به این بانک‌ها نیز فراهم شده‌است. نمونه‌ای از خروجی برنامه در شکل ۵ نشان داده شده‌است.

شکل ۵. نمونه‌ای از خروجی برنامه

۶. تجزیه و تحلیل داده‌ها

۱،۶. آنالیز بانک اطلاعاتی یک استان

پس از مراحل طراحی و پیاده‌سازی، که قبلاً بیان‌شد، در ابتدا از بانک بیماری‌های دانشگاه علوم پزشکی و خدمات درمانی استان زنجان، با انواع مختلف داده، برای ارزیابی صحت و دقت برنامه، استفاده‌شد. بانک‌های اطلاعاتی، مربوط به سال ۱۳۹۳ استان زنجان است و حاوی اطلاعات افرادی است که به‌علت بیماری در استان زنجان فوت‌شده‌اند. این اطلاعات هرساله توسط دانشگاه علوم پزشکی جمع‌آوری‌شده تا بعد از اصلاح موارد نادرست و ثبت‌نشده و در نهایت حذف رکوردهای تکراری، در تحلیل‌های مدیریتی مورد استفاده قرارگیرد. پیش‌پردازش این اطلاعات هرساله ، دستی انجام‌می‌شد که زمان‌بر و دارای خطای انسانی است.

پس از فرآیند ادغام بانک اطلاعاتی شهرستان‌های استان زنجان (هشت شهرستان: ابهر، ایجرود، خدابنده، خرمدره، زنجان، طارم، ماهنشان) ۵۴۶۴ رکورد در بانک اطلاعاتی ill ایجادشد.

مرحله بعد، اجرای فرآیند خطایابی برروی بانک اطلاعاتی بود که خطاهای موجود در بانک شناسایی‌شد. به منظور یافتن خطاهای موجود در بانک اطلاعاتی از استورپراسیجرهای اس‌کیوال استفاده‌شد بنابراین زمان اجرای الگوریتم جهت یافتن خطاهای بانک اطلاعاتی به کمتر از سه ثانیه وقت نیاز دارد. نتایج خاصل از این خطایابی در جدول ۱ برای هر فیلد اطلاعاتی این بانک نشان‌داده شده‌است.

جدول ۱: تعداد خطا در فیلدهای بانک اطلاعاتی

فیلد	میزان خطا قبل از اصلاح	درصد خطا قبل از اصلاح	درصد خطا بعد از اصلاح
جنسیت	۳۱	٪۰,۵۶	٪۰
تاریخ تولد	۸	٪۰,۱۴	٪۰
تاریخ فوت	۱۳	٪۰,۲۳	٪۰
نوع زندگی	۸۰	٪۱,۴۶	٪۰
بیماری	۸۷۳	٪۱۵,۹۸	٪۰
کدICD	۳۲۶	٪۵,۹۷	٪۰
استان محل تولد	۱۰	٪۰,۱۸	٪۰
شهرستان محل تولد	۳۸۹	٪۷,۱۲	٪۰
استان محل زندگی	۲۰	٪۰,۳۶	٪۰
شهرستان محل زندگی	۵	٪۰,۰۹	٪۰
استان محل فوت	۲	٪۰,۰۴	٪۰
شهرستان محل فوت	۳	٪۰,۰۵	٪۰
محل فوت	۱۵	٪۰,۲۷	٪۰
ثبت کننده	۳۴	٪۰,۶۲	٪۰
منبع اعلام کننده	۲۳	٪۰,۴۲	٪۰
رکورد های تکراری	۶۷۳	٪۱۲	٪۰,۰۲

از این مقادیر به‌دست آمده می‌توان به این نتیجه رسید که درصد کل خطای این بانک اطلاعاتی با ۵۴۶۴ رکورد ۲,۸۴٪ می‌باشد. به این معنی است که ممکن است یک رکورد، در این بانک اطلاعاتی فاقد خطا بوده در حالی که ممکن است رکورد دیگری در اکثر فیلدهای آن خطایی رخ داده باشد. به منظور افزایش سرعت در فرآیند یافتن خطاها در بانک اطلاعاتی از استور پراسیجرهای اس‌کیوال استفاده‌شد لذا زمان اجرای جهت یافتن خطاهای موجود در این بانک سه ثانیه به‌دست‌آمد.

بعد از یافتن داده‌های آلوده، فیلد رکوردهایی که دارای داده آلوده بود باید اصلاح‌می‌شد. لذا فرآیند اصلاح برروی رکوردهایی که دارای داده آلوده بود اجراشد.

رکوردهایی که در فیلدهای جنسیت، بیماری، کدICD، استان و شهرستان محل تولد، محل زندگی و محل فوت دارای داده آلوده بود از طریق استورپراسیجرهای طراحی‌شده در اس‌کیو ال و با استفاده از فرهنگ لغت طراحی شده برای آن فیلد اصلاح گردید. و رکوردهایی که در فیلدهای تاریخ تولد و فوت، نوع زندگی، محل فوت، ثبت کننده، منبع اعلام کننده دارای داده آلوده بود از طریق الگوریتم‌هایی که در سی‌شارپ طراحی و در فصل قبل به آن اشاره شد اصلاح گردید.

در نهایت حذف رکوردهای تکراری با اجرای استور پراسیجرهایی که در محیط اس‌کیوال طراحی شده بود، برروی داده‌ها بانک اعمال گردید. مدت زمان اجرای فرآیند اصلاح برای بانک اطلاعاتی با ۵۴۶۴ رکورد ۶۳ ثانیه بود.

در آخرین مرحله تست، مجددا بانک اطلاعاتی اصلاح شده را با روش-های طراحی شده در برنامه مورد بررسی قرار دادیم مشاهد شد که تعداد رکوردها از ۵۴۶۴ به ۴۷۹۱ رکورد کاهش پیدا کرد. بدین معنی است که یک شخص در دو شهرستان مختلف ثبت شده و یا کاربر یک رکورد را بیش از یک بار ثبت نموده است. با مقایسه نتایج به‌دست آمده (مطابق شکل ۴۱) مشاهده می‌گرددکه میزان خطا در ۵۴۶۴ رکورد ۲,۸۴٪ به ۰,۰۲٪ در ۴۷۹۱ رکورد کاهش یافت که این خطا مربوط به رکوردی است که دارای کدملی یکسان ولی با نام، نام پدر، محل فوت متفاوت، اطلاعاتش ثبت گردیده است.که با پیگیری قابل اصلاح می‌باشد. همچنین مدت زمان اجرای برنامه جهت‌یافتن خطا و اصلاح آن ۶۵ ثانیه به‌دست‌آمد.

با اجرای فرآیند اصلاح داده‌های آلوده بانک اطلاعاتی با خطای ۰,۰۲٪ اصلاح شد. که این خطا مربوط به رکوردی است که دارای کدملی یکسان ولی با نام، نام پدر، محل فوت متفاوت، اطلاعاتش ثبت‌گردیده است بنابراین با گروه‌بندی که در فرآیند حذف رکوردهای تکراری انجام شده‌است این رکورد در هیچ یک از گروه‌های قرارنگرفته لذا باعث ایجاد خطا در سیستم ما گردیده‌است.

۲,۶. آنالیز بانک اطلاعاتی دو استان

به منظور بررسی مجدد معماری طراحی‌شده، این بار تعداد رکوردهای بانک اطلاعاتی را با ادغام اطلاعات دو استان زنجان و ایلام تست نمودیم. با ادغام این دو بانک اطلاعاتی ۱۰۲۰۳ رکورد ایجادشد. مرحله بعد اجرای الگوریتم خطایابی برروی این بانک اطلاعاتی است که نتایج آن در جدول ۲ نشان داده شده‌است.

جدول۲. میزان خطا در بانک اطلاعاتی ایجاد شده از ادغام دو استان زنجان و ایلام

فیلد	تعداد قبل از اصلاح	درصد خطاها قبل از اصلاح	درصد خطاها بعد از اصلاح
جنسیت	۱۰۸	٪۱,۰۵	٪۰
تاریخ تولد	۸	٪۰,۰۸	٪۰
تاریخ فوت	۱۳	٪۰,۱۳	٪۰
نوع زندگی	۴۶۰	٪۵,۲۹	٪۰
بیماری	۱۹۲۵	٪۱۸,۸۷	٪۰
ICDکد	۱۰۷۰	٪۱۰,۴۹	٪۰
استان محل تولد	۱۰	٪۰,۱	٪۰
شهرستان محل تولد	۵۶۵	٪۵,۵۳	٪۰
استان محل زندگی	۲۰	٪۰,۲	٪۰
شهرستان محل زندگی	۵	٪۰,۰۵	٪۰
استان محل فوت	۵۲	٪۰,۵۱	٪۰
شهرستان محل فوت	۳	٪۰,۰۳	٪۰
محل فوت	۱۵	٪۰,۱۵	٪۰
ثبت کننده	۳۴	٪۰,۳۳	٪۰
منبع اعلام کننده	۲۳	٪۰,۲۲	٪۰
رکورد های تکراری	۲۲۹۰	٪۲۲,۴۴	٪۰,۰۱

میزان کل خطای بانک‌اطلاعاتی بعد از ترکیب دو استان برابر ۴,۰۹٪ می‌باشد. بعد از خطایابی نوبت به اصلاح اطلاعات، حذف رکوردهای تکراری، محاسبه سن و گروه‌های سنی است. بانک اطلاعاتی بعد از تکرارگیری دارای ۷۹۱۳ رکورد شد. و نتایج به‌دست آمده بعد از اجرای الگوریتم در جدول ۲ نشان داده شده است.

در بانک اطلاعاتی رکوردی وجود دارد که دارای کدملی یکسان ولی با نام، نام پدر، محل فوت متفاوت، اطلاعاتش ثبت‌گردیده‌است بنابراین با گروه‌بندی که در فرآیند حذف رکوردهای تکراری انجام‌شده‌است این رکورد در هیچ یک از گروه‌های قرارنگرفته لذا باعث ایجاد خطا در سیستم ما گردیده‌است.

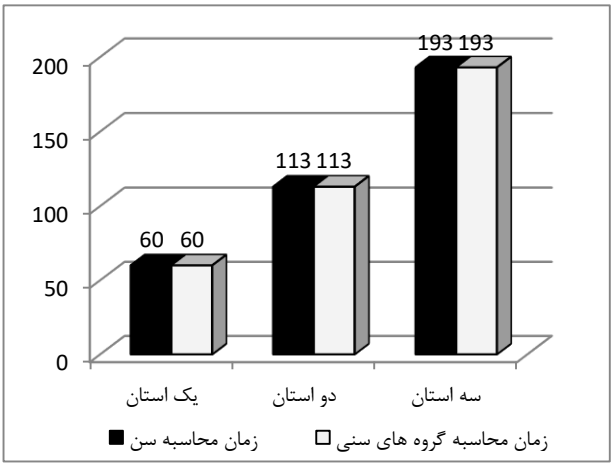
۳,۶. آنالیز بانک اطلاعاتی سه استان

در مرحله سوم تست با اضافه کردن استان همدان، تعداد رکوردهای بانک اطلاعاتی را به تعداد ۱۴۷۳۹ رکورد افزایش‌دادیم. میزان کل خطای بانک اطلاعاتی بعد از ترکیب سه استان برابر ۳,۷٪ شد. بعد از خطایابی نوبت به اصلاح اطلاعات، حذف رکوردهای تکراری، محاسبه سن و گروه‌های سنی است. بانک اطلاعاتی بعد از تکرارگیری دارای ۱۲۴۲۴ رکورد شد. و نتایج حاصل به‌دست آمده بعد از اجرای الگوریتم در جدول ۳ نشان داده‌شده‌است.

جدول ۳: میزان خطا در بانک اطلاعاتی ادغامی سه استان زنجان، همدان و ایلام

فیلد	تعداد قبل از اصلاح	درصد خطاها قبل از اصلاح	درصد خطاها بعد از اصلاح
جنسیت	۱۶۷	٪۱,۱۳	٪۰
تاریخ تولد	۸	٪۰,۰۵	٪۰
تاریخ فوت	۱۳	٪۰,۰۹	٪۰
نوع زندگی	۵۳۹	٪۳,۵۶	٪۰
بیماری	۲۹۵۶	٪۲۰,۰۵	٪۰
ICDکد	۱۸۳۱	٪۱۲,۴۲	٪۰
استان محل تولد	۱۰	٪۰,۰۷	٪۰
شهرستان محل تولد	۷۶۵	٪۵,۱۹	٪۰
استان محل زندگی	۲۲	٪۰,۱۵	٪۰
شهرستان محل زندگی	۵	٪۰,۰۳	٪۰
استان محل فوت	۵۲	٪۰,۳۵	٪۰
شهرستان محل فوت	۳	٪۰,۰۲	٪۰
محل فوت	۱۵	٪۰,۱۰	٪۰
ثبت کننده	۳۴	٪۰,۲۳	٪۰
منبع اعلام کننده	۲۳	٪۰,۱۶	٪۰
رکورد های تکراری	۲۳۱۰	٪۱۵,۶۷	٪۰,۰۰۸

در این معماری سن و گروه‌های سنی نیز محاسبه می‌گردد با افزایش تعداد رکوردها مدت زمان محاسبه در قسمت سن و گروه‌های سنی افزایش پیدا می‌کند به‌گونه‌ای که برای محاسبه سن و گروه‌های سنی در بانک اطلاعاتی یک استان (استان زنجان)، مدت زمان محاسبه هرکدام ۶۰ ثانیه، که در مجموع ۱۲۰ ثانیه و در ترکیب دو استان (استان زنجان و ایلام) مدت زمان محاسبه هرکدام ۱۱۳ ثانیه، که در مجموع ۲۲۶ ثانیه به طول انجامید. حال با ترکیب سه استان زنجان، ایلام و همدان که تعداد کل رکوردهای این سه استان، بعد از حذف موارد تکراری ۱۲۴۲۴ رکوردشد، مدت زمان محاسبه سن و گروه‌های سنی برای هرکدام، ۱۹۳ ثانیه که در مجموع، ۳۸۶ ثانیه به‌دست‌آمد. (شکل ۶)

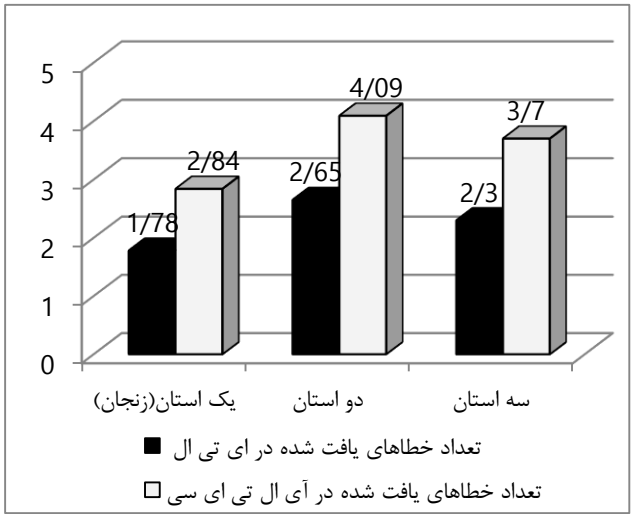


شکل ۶: نمودار مقایسه زمان محاسبه سن و گروه های سنی برای یک ، دو و سه استان (به ثانیه)

با توجه به تست‌های انجام‌شده در سه مرحله قبل می‌توان نتیجه‌گرفت که با افزایش تعداد رکوردهای بانک اطلاعاتی (ادغام بانک اطلاعاتی استان‌های دیگر) سرعت یافتن خطاهای و اصلاح رکوردهایی که دارای داده‌های آلوده بودند تفاوتی ندارد زیرا در به‌روزرسانی بانک اطلاعاتی از استورپراسیجرهای اس‌کیوال استفاده‌شده که مستقیم با بانک اطلاعاتی کارمی‌کند و زمان به‌روزرسانی را کاهش‌می‌دهد. همچنین نشان‌دادیم زمان محاسبه سن و گروه‌های سنی با افزایش رکوردهای بانک اطلاعاتی افزایش‌می‌یابد.

۴,۶. مقایسه ای‌تی‌ال با مدل پیشنهادی

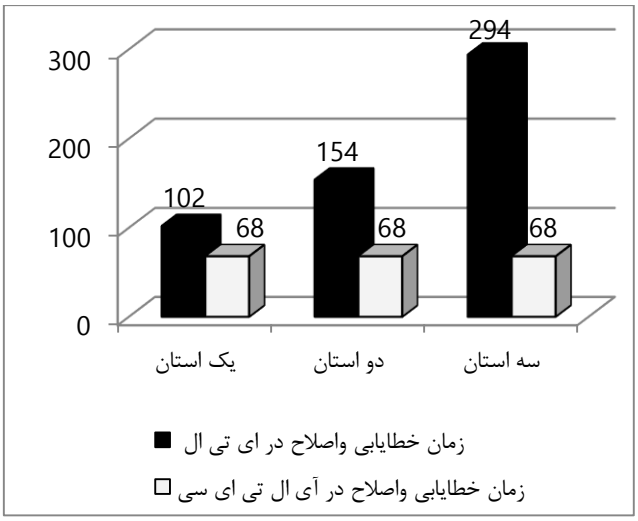
در انتها به منظور نشان‌دادن عملکرد برنامه طراحی‌شده، مقایسه‌ای بین معماری آی‌ال‌تی‌ای‌سی و ای‌تی‌ال انجام‌گرفت. لذا برای همین منظور داده‌هایی که در معماری آی‌ال‌تی‌ای‌سی استفاده‌کرده‌ایم را در معماری ای‌تی‌ال استفاده‌نمودیم و با توجه به موارد ذیل معماری آی‌ال‌تی‌ای‌سی دارای عملکرد بهتری نسبت به ای‌تی‌ال شد:



شکل ۷.:مقایسه تعداد کل خطاها در دو مدل ای‌تی‌ال و آی‌ال‌تی‌ای-سی برای یک، دو و سه استان (قبل از فرایند اصلاح)

فرآیند خطایابی و اصلاح بر روی حداکثر دو نوع داده مختلف در معماری ای‌تی‌ال انجام‌می‌شود لذا خطاهای کمتری نسبت به معماری ارائه شده (آی‌ال‌تی‌ای‌سی)، به‌دست‌آمد و در نتیجه اصلاح کمتری نسبت به معماری آی‌ال‌تی‌ای‌سی انجام‌گرفت. (شکل ۷)

به دلیل استفاده از محیط‌های عملیاتی موقت، ابتدا اطلاعات به این محیط منتقل، و فرآیند اصلاح و خطایابی انجام‌شده و از این قسمت اطلاعات به محل اصلی منتقل، تا از محل اصلی به لایه بعدی منتقل‌شود. که سبب‌می‌شود زمان اجرای معماری ای‌تی‌ای بیشتر از معماری آی‌ال-تی‌ای‌سی شود(شکل ۸)



شکل ۸.: نمودار مقایسه زمان اجرا (خطایابی واصلاح) در دو معماری ای‌تی‌ال و آی‌ال‌تی‌ای‌سی برای یک،دو و سه استان

بعد از فرآیند خطایابی و اصلاح بانک‌های‌اطلاعاتی، بانک‌اطلاعاتی اصلاحی مجدد توسط برنامه طراحی‌شده ارزیابی‌می‌شود تا میزان اصلاح داده‌های آلوده بررسی‌شود در معماری آی‌ال‌تی‌ای‌سی و ای‌تی‌ال با ۱۲۴۲۴ رکورد، دارای خطای ۰,۰۰۸٪ بود که این خطا به روش پیاده-سازی مربوط نبوده و خطای کاربر در زمان ورود اطلاعات بوده که یک کدملی برای دو فرد مختلف ثبت شده‌است. ولی در معماری ای‌تی‌ال در مقایسه با معماری آی‌ال‌تی‌ای‌سی، هنوز داده‌های وجوددارد که در مرحله خطایابی یافت‌نشده تا اصلاح‌شوند.

۷. نتیجه‌گیری

جهت تست معماری ای‌تی‌ال، از بانک‌اطلاعاتی با ۵۴۶۴ رکورد (بیماری-های استان زنجان)، ۱۰۲۰۳ رکورد (بیماری‌های دو استان زنجان و ایلام) و ۱۴۷۳۹ رکورد (بیماری‌های سه استان زنجان، ایلام و همدان) استفاده شد (در معماری آی‌ال‌تی‌ای‌سی استفاده شده بود). اجرای فرآیند خطایابی و اصلاح برنامه با بانک اطلاعاتی که دارای ۵۴۶۴ رکورد بود ۱۰۲ ثانیه زمان برد و ۱,۷۸٪ میزان خطاها در بانک‌اطلاعاتی شد. جهت تست دوباره، از بانک اطلاعاتی که دارای ۱۰۲۰۳ رکورد بود، استفاده مودیم این بار زمان اجرای فرایند خطایابی و اصلاح برابر ۱۵۴ ثانیه، و تعداد خطاهای برابر ۲,۶۵ ٪ بود. در نهایت از بانک‌اطلاعاتی که دارای ۱۴۷۳۹ رکورد بود، استفاده‌کردیم که زمان اجرای آن ۲۹۳ ثانیه و تعداد خطاها در فرآیند خطایابی و اصلاح برابر ۲,۳٪ بود.

با توجه به آزمایش‌های انجام شده برروی سه بانک اطلاعاتی به گونه-ای که هر بار رکوردهای آن را افزایش‌دادیم به این نتیجه رسیدیم که معماری طراحی‌شده دارای دقت و سرعت بالاتری دریافتن خطا و اصلاح آن‌ها نسبت معماری ای‌تی‌ال است. به گونه‌ای که معماری ارائه‌شده تعداد خطاهای بیشتری یافته و اصلاح می‌نماید و در نهایت می‌توان اذعان-داشت که این معماری در فرآیند اصلاح ۱۲۴۲۴ رکورد ۰,۰۰۸٪ خطا دارد که این خطا مربوط به رکوردی است که دارای کدملی یکسان ولی با نام، نام پدر، محل فوت متفاوت، اطلاعاتش ثبت گردیده‌است بنابراین با گروه‌بندی که در فرآیند حذف رکوردهای تکراری انجام‌شده‌است این رکورد در هیچ‌یک از گروه‌ها قرارنگرفته لذا باعث ایجاد خطا در سیستم ما گردیده‌است.

مراجع

[1] Al Farsi, Budour Ahmed; Saini, Dinesh Kumar. Business Intelligence Design Model (BIDM) for University. International Journal of Computer Applications, Vol. 111 , No 14, (2015),pgs.43-49

[2]Ghazanfari.M,M.Jafari.M, Rouhani.S.A tool to evaluate the business intelligence of enterprise systems.Scientia Iranica Transactions E(Industrial Engineering), Vol.16, No.6, (2011), pgs.1579–1590

[3] Golfarelli , Matte.New Trends in Business Intelligence.Conference: Proceedings of the 28th International Convention MIPRO (BIS&DE&ISS), MIPRO 2005,(May 30-June 03, 2005), Opatija, Croatia, https://www.researchgate.net/publication/ 221535705_New_Trends_in_Business_Intelligence , 2016/29/01

[4] Rupali Gil,R; Singh,J.A Review of Contemporary Data Quality Issues in Data Warehouse ETL Environment. Journal on Today's Ideas Tomorrow's Technologies,Vol. 2, No. 2, (2014), pgs.153-160

[5] Ong1, In Lih; Siew, Pei Hwa; Wong, Siew Fan.A Five-Layered Business Intelligence Architecture.IBIMA Publishing,Vol.2011, http://www.ibimapublishing .com/journals/CIBIMA/cibima.html,2016/29/01

[6] Kalelkar, Medha; Churi, Prathamesh; Kalelkar, Deepa. Implementation of Model-View-Controller Architecture Pattern for Business Intelligence Architecture. International Journal of Computer Applications, Vol.102, No.12, (2014), pgs.16-21

[7] Anand , Nitin. ETL and its impact on Business Intelligence. International Journal of Scientific and Research Publications, Vol. 4, Issue 2, (2014), pgs.1-3

[8] Tank, Darshan M; Ganatra, Amit; Kosta, Y P. Speeding ETL Processing in Data Warehouses Using High-Performance Joins For Changed Data Capture (CDC). International Conference on Advances in Recent Tecnologic in Communication and Computing, (2010) International Conference on Source: IEEE Xplore, pgs.365-368

[9] Rajashree, Y.Patil; Kulkarni,R.V. A Review of Data Cleaning Algorithms for Data Warehouse Systems. International Journal of Computer Science and Information Technologies, Vol. 3 , No.5 ,(2012), pgs.5212 – 5214

[10] Kirange, Mayuri; Makhijani, R.K. Revolution In DW By Solving Causes Of Data Quality Problems In DW And ETL. International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 1, (2015), pgs.64 – 73

[11] Devi1,S; Kalia, A. Study of Data Cleaning & Comparison of Data Cleaning Tools. International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 3,(2015), pgs.360 – 370

[12] Kabiri,A.; Chiadmi,D. Survey on ETL processes. Journal of Theoretical and applied Information Technology. Vol. 54, No. 2, (2013), pgs.219-229

[13] Talebzadeh,Hossein. A Service-Based Framework for ETL Process Based on Metadata. Journal of Basic and Applied Scientific Research,Vol.2, No.1, (2012), pgs.54-59

[14] Ganapavarapu, Vinaya Bharadwaj. "Designing and Implementing a Data Warehouse using Dimensional Modeling" Master of Science, Computer Engineering , The University of New Mexico, Heileman,Gregory ,(2014)

[15] NEDELCU, Bogdan.Business Intelligence Systems. Database Systems Journal Vol. IV, No. 4, (2013), pgs.12-20

[16] Vora, Mital; Vora, Jelam; Jani, Dr. N. N. Modelling Architecture for Multimedia Data Warehouse. International Journal of Innovative Research in Science Engineering and Technology, Vol. 4, Issue 1, January (2015) ,pgs.18699-18703

[17] Kushanoor, Akbar; Dr. Krishna, S.Murali; Sagar Reddy, T. Vidya. ETL Process Modeling In DWH Using Enhanced Quality Techniques. International Journal of Database Theory and Application Vol. 6, No. 4, August,(2013),pgs.179-198

[18] Dubey, Alok ; Kamal, Archana ; Gupta, Suresh C. Effects of Aggregation and Data Size on Query Performance and Memory Requirements of a Data Warehouse. International Conference on Reliability Optimization and Information Technology , India,(2014), IEEE Xplore, pgs.(99-104)

[19] Khalid,Muhammad[et al].Challenges of Dimensional Modeling in Business Intelligence Systems. International Journal of Computer & Organization Trends, Vol. 21, No.1, (2015),pgs.14-15

[20] Ifeanyi, Nwakanma [et al].The Role of Data Warehousing Concept for Improved Organizations Performance and Decision Making . International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 10, (2014), pgs.451 – 455

[21] Ali El-Sappagh, Shaker H.; Ahmed Hendawi, Abdeltawab M.; El Bastawissy, Ali Hamed. A proposed model for data warehouse ETL processes. Journal of King Saud University, Computer and Information Sciences (2011) 23, pgs.91–104

[22] Porwa, Sonal; Vora, Deepali. A Comparative Analysis of Data Cleaning Approaches to Dirty Data. International Journal of Computer Applications ,Volume 62, No.17, (2013),pgs.30-34

[23]Bhattacharjee, Arup Kumar; Mallick ,Atanu; Dey, Arnab; Bandyopadhyay, Sananda. Data Cleaning in Text File. IOSR Journal of Computer Engineering, Volume 9, Issue 2 (2013), Pgs. 17-21

[24] Bhattacharjee, Arup Kumar; Mallick ,Atanu; Dey, Arnab ; Bandyopadhyay , Sananda. Enhanced Technique for Data Cleaning in Text File. International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, (2013),pgs.229-233

[25] Khedri, Ridha; Chiang, Fei; SabriAn, Khair Eddin. Algebraic Approach Towards Data Cleaning. The 4th International Conference on Emerging Ubiquitous Systems and Pervasive Networks , Procedia Computer Science 21 (2013), www.sciencedirect.com, pgs.50 – 59

[26] Sheoran, Jyoti. Issues of Data Quality in Data Warehouses. International Conference on Advances in Computer Engineering & Applications (ICACEA-2014) at IMSEC,GZB, pgs.6-8, http://www.ijcaonline.org/proceedings/icacea/ number6/15835-1465 ,2016/01/29

[27] Save, Ashwini M.; Kolkur , Seema. Hybrid Technique for Data Cleaning. International Journal of Computer Applications Proceedings on National Conference on Role of Engineers in National Building NCRENB,(2014), pgs. 4-8

[28] Kulkarni, Prerna S; Bakal, J.W.Hybrid Approaches for Data Cleaning in Data Warehouse. International Journal of Computer Applications, Vol.88 , No.18, (2014), pgs.7-10

[29] Bhattacharjee, Arup Kumar; Chatterjee, Partha; Shaw, Mukesh Prasad ; Chakraborty, Manomoy. ETL based Cleaning on Database. International Journal of Computer Applications ,Vol 105, No. 8,(2014), pgs.34-40.

پی‌نوشت

1. Howard Dresner
2. Ghoshal
3. Kim
4. Eckerson
5. OLAP: Online Analytical Processing
6. Data Mining
7. Data Source
8. ETL(Extraction-Transformation-Loading)
9. Extraction
10. Transformation
11. Loading
12. Data Warehouse
13. Mata Data
14. End user
15. Subject Oriented
16. Time variant
17. Non volatile
18. Dimensional Modeling
19. Fact Tables
20. Dimensional Tables
21. Star Schema
22. Snowflake Schema
23. Visualization Tools
24. BPM:Business Performance Management
25. Join
26. Aggregation
27. EMD:Entity Mapping Diagram
28. Conceptual model
29. Stored Procedure
30. ILTEC:Integrated Load Transformation Extraction Compact