Journal of Information and Communication Technology

Volume 16, Issue 59-60, Spring and Summer 2024, pp. 1-16

Providing a New Solution in Selecting Suitable Databases for Storing Big Data in the National Information Network

Mohammad Reza Ahmadi[1], Davood Maleki[2], Ehsan Arianyan[3*]

[1]Associate Professor, ICT Research Institute, Tehran, Iran
[2]Faculty Member, ICT Research Institute, Tehran, Iran
[3]Assistant Professor of ICT Research Institute, Tehran, Iran

Received: 31 October 2021, Revised: 02 May 2022, Accepted: 14 August 2022
Paper type: Research

Abstract

The development of infrastructure and applications, especially public services in the form of cloud computing, traditional models of database services and their storage methods have faced sever limitations and challenges. The increasing development of data service productive tools and the need to store the results of large-scale processing resulting from various activities in the national network of information and data produced by the private sector and pervasive social networks has made the process of migrating to new databases with appropriate features inevitable. With the expansion and change in the size and composition of data and the formation of big data, traditional practices and patterns do not meet new needs. Therefore, it is necessary to use data storage systems in new and scalable formats and models. This paper reviews the essential solution regarding the structural dimensions and different functions of traditional databases and modern storage systems and technical solutions for migrating from traditional databases to modern ones suitable for big data. Also, the basic features regarding the connection of traditional and modern databases for storing and processing data obtained from the national information network are presented and the parameters and capabilities of databases in the standard platform context and Hadoop context are examined. As a practical example, a combination of traditional and modern databases using the balanced scorecard method is presented as well as evaluated and compared.

Keywords: Big Data, Traditional Databases, Big Data Databases, Databases Selection.

[*] Corresponding Author's email: ehsan_arianyan@itrc.ac.ir

دوفصلنامه
فناوری اطلاعات و ارتباطات ایران

سال شانزدهم، شماره های ۵۹ و ۶۰، بهار و تابستان ۱۴۰۳، صفحه ۱ الی ۱۶

راهکاری نوین در انتخاب پایگاه‌های داده مناسب جهت ذخیره‌سازی کلان داده‌ها در شبکه ملی اطلاعات

محمدرضا احمدی[۱]، داود ملکی[۲]، احسان آریانیان[۳*]

[۱] دانشیار پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

[۲] عضو هیات علمی پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

[۳] استادیار پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

تاریخ دریافت: ۱۴۰۰/۰۸/۰۹ تاریخ بازبینی: ۱۴۰۱/۰۲/۱۲ تاریخ پذیرش: ۱۴۰۱/۰۵/۲۳

نوع مقاله: پژوهشی

چکیده

توسعه زیرساخت‌ها و برنامه‌های کاربردی به خصوص سرویس‌های همگانی در قالب رایانش ابری، الگوهای سنتی خدمات پایگاه‌های داده و روش‌های ذخیره‌سازی آنها را با محدودیت‌ها و چالش‌های جدی روبرو ساخته است. توسعه روزافزون ابزارهای مولد خدمات داده‌ای و لزوم ذخیره‌سازی نتایج پردازش‌های بزرگ وگسترده حاصل از فعالیت‌های مختلف در شبکه ملی اطلاعات و داده‌های تولیدی بخش خصوصی و شبکه‌های فراگیر اجتماعی، روند مهاجرت به پایگاه‌های داده نوین با ویژگی‌های مناسب را اجتناب‌ناپذیر کرده است. با گسترش و تغییر حجم و ترکیب داده‌ها و شکل‌گیری کلان داده‌ها، عملکردها و الگوهای سنتی پاسخگوی نیازهای جدید نیستند. بنابراین لزوم استفاده از سیستم‌های ذخیره‌سازی داده در قالب‌ها و مدل‌های نوین و مقیاس‌پذیر را ضروری ساخته است. در این مقاله راهکارهای اساسی در خصوص ابعاد ساختاری و کارکردهای مختلف پایگاه‌های داده سنتی و سیستم‌های ذخیره‌سازی نوین بررسی گردیده و راهکارهای فنی جهت مهاجرت از پایگاه‌های داده سنتی به نوین و مناسب برای کلان داده‌ها ارائه می‌گردد. همچنین ویژگی‌های اساسی در خصوص پیوند پایگاه‌های داده سنتی و نوین جهت ذخیره و پردازش داده‌های حاصل از شبکه ملی اطلاعات ارائه شده و پارامترها و قابلیت‌های پایگاه‌های داده در بستر استاندارد پایه و بستر هدوپ بررسی شده است. به عنوان یک نمونه عملیاتی یک راهکار ترکیبی از پایگاه داده سنتی و نوین با استفاده از روش کارت امتیازی متوازن ارائه شده و مورد ارزیابی و مقایسه قرار گرفته است.

کلیدواژگان: کلان داده‌ها، پایگاه‌های داده سنتی، پایگاه داده‌ی کلان داده، انتخاب پایگاه‌های داده.

* رایانامهٔ نویسنده مسؤول: ehsan_arianyan@itrc.ac.ir

۱- مقدمه

اولین سیستم مدیریت پایگاه داده برای اولین بار توسط یکی از پیشگامان این شاخه به نام چارلز بکمن[1] گسترش یافت[۱]. مقالات بکمن این را نشان داد که فرضیات او کارایی بسیار مؤثرتری برای دسترسی به وسایل ذخیره‌سازی را مهیا می‌کند. به‌دنبال آن مدل رابطه‌ای به همراه عملگرهای جبری و اصول آن در دهه هفتاد میلادی توسط ادگار کاد[2] ارائه شدند[۱]. کاد به دلیل این ابتکار، جایزه تورینگ را از آن خود ساخت که معادل جایزه نوبل در علم رایانه است. او مدل‌های موجود را مورد انتقاد قرار می‌داد. برای مدتی نسبتاً طولانی این مدل در مجامع علمی مورد تأیید قرار گرفت. اولین محصول موفق برای میکروکامپیوترها dBASE بود که برای سیستم‌عامل‌های CP/M و PC-DOS/MS-DOS ساخته شد. در دهه ۱۹۸۰ پژوهش بر روی پایگاه‌های مدل توزیع شده و ماشین‌های پایگاهی متمرکز شد، اما تأثیر کمی بر بازار گذاشت. در سال ۱۹۹۰ توجه متخصصان به مدل شیءگرا جلب شد. این مدل جهت کنترل داده‌های مرکب لازم بود و به خوبی با استفاده از پایگاه‌های داده غیرسنتی، در کاربردهایی نظیر مهندسی داده (شامل مهندسی نرم‌افزار منابع) و مهندسی داده‌های چند رسانه‌ای کار می‌کرد. در سال ۲۰۰۰ نوآوری تازه‌ای رخ داد و پایگاه XML به وجود آمد. هدف این مدل از بین بردن تفاوت بین مستندات و داده‌ها بود و کمک می‌کرد که منابع اطلاعاتی چه ساخت یافته و چه غیرساخت‌یافته در کنار هم قرار گیرند. مفاهیم پایگاه‌های داده سنتی رابطه‌ای به همراه زبان برنامه‌نویسی SQL، مجموعه‌ای از جدول‌ها، ستون‌ها و سطرهای داده‌ای و روابط جدول‌ها می‌باشند. اما، برخی محدودیت‌ها و مشکلات باعث مطرح شدن مفاهیم جدیدی در پایگاه‌های داده شده است. مشکلات اصلی مدل رابطه‌ای سنتی عبارتند از:

- محدودیت فضا و ناکارآمدی فناوری‌های سنتی ذخیره‌سازی برای حجم عظیم اطلاعات، که چالش اصلی آن‌را باید در نحوه جستجوی داده‌ها و یافتن ارتباط‌ها و نظم‌های پنهان در میان حجم عظیم داده جست‌وجو کرد.
- در محیط جدید کلان داده‌ها[3] داده‌هایی از انواع گوناگون نظیر متن، تصویر، فیلم، صوت و انواع دیگر داده‌ها وجود دارد. ابزارهای مدیریتی پایگاه‌های داده سنتی قادر به جمع‌آوری، ذخیره‌سازی، جست‌وجو، اشتراک‌گذاری، تحلیل و نمایش همزمان تمام این انواع داده‌ای به صورت کارا نیستند.
- یکی از مهم‌ترین مسائل مرتبط با کلان داده‌ها، مشکل بودن کار با آنها به‌وسیله پایگاه‌داده‌ای رابطه‌ای و بسته‌های نرم‌افزاری مرتبط با آن است. چرا که برای پردازش این داده‌ها در یک زمان معقول به نرم‌افزارهای به شدت موازی شده با قابلیت اجرا بر روی تعداد زیادی سرور نیاز است، این در حالی است که ساختار سنتی پایگاه‌های داده رابطه‌ای برای اجرای برنامه‌های موازی کارا نیستند.

از طرف دیگر مدل رابطه‌ای دارای محدودیت‌های زیادی می‌باشد که در ادامه به مواردی از آنها اشاره خواهیم کرد:

- کارایی پایین برای داده‌های با توالی نوشتن بالا و توالی خواندن کم- نظیر شمارنده‌های بازدید صفحات وب و دستگاه‌های وقایع نگار فضایی.
- کارایی پایین برای داده‌های با توالی خواندن بالا و توالی نوشتن بسیار کم- نظیر داده‌های گذرا و تصاویر، اسناد و پردازش تصویر در HTML با دسترسی تکراری.
- کارایی پایین برای کاربردهایی با نیازمندی‌های در دسترس بودن بالا[4] و با توقف خدمات[5] بسیار کم- این مورد در مدل رابطه‌ای با کمبودهایی همراه است و به خوبی از عهده آن بر نمی‌آید. این نوع کاربردها بیش از هر چیز به مقیاس‌پذیری افقی و امکان توسعه بر روی ماشین‌های مختلف شبکه نیاز دارند.
- کارایی پایین برای داده‌هایی که باید در نقاط مختلف جغرافیایی با هم همگام‌سازی شوند- استفاده از پایگاه‌های داده سنتی برای داده‌های عظیمی که در خوشه‌های مختلف یک شبکه بزرگ سازمانی با دفاتر پراکنده در سطح جغرافیایی وسیع ذخیره شده‌اند و نیاز است تا همواره با بالاترین سرعت و کمترین هزینه ممکن با هم همگام‌سازی شوند، کارا و قابل استفاده نیست و هزینه‌های بسیاری را در برخواهد داشت.
- کارایی پایین برای داده‌های بزرگ تجاری یا مرتبط با تحلیل وب که طرح[6] خاصی ندارند- این داده‌ها تقریباً و قالب از پیش‌تعیین شده‌ای ندارند و بر اساس محتوای متغیر موجود بر روی وب تولید می‌شوند و در بیشتر موارد به فعالیت کاربران و سیستم‌های نرم‌افزاری مرتبط و وابسته هستند. اغلب نیاز است تا چنین داده‌هایی به صورت موازی ذخیره‌سازی شوند و امکانات پرس‌وجوپذیری غنی مهیا باشد تا به‌خوبی قابل تحلیل

[1] Charles Bachman
[2] Edgar Frank Codd
[3] Big Data
[4] High Availability
[5] Downtime
[6] Schema

باشند که در مدل‌های رابطه‌ای سنتی این قابلیت وجود ندارد.

کارلو استروزی[1] نخستین بار در سال ۱۹۹۸ عبارت NoSQL را برای اشاره به پایگاه‌های داده‌ای سبک و متن باز رابطه‌ای به کار گرفت که از مدل رابطه‌ای SQL استفاده نمی‌کردند[۲]. هر چند بعدها وی به این نکته اشاره کرد که این عبارت و مفهوم پشت آن، کاملاً از مدل رابطه‌ای جداشده و بهتر است آن را غیررابطه‌ای یا NoREL[2] بنامیم. عبارت NoSQL مفهومی است که برای مشخص کردن پایگاه‌های داده‌ای به کار می‌رود که به شدت با پایگاه‌های داده‌ای رابطه‌ای سنتی متفاوت هستند. این پایگاه‌های داده اغلب با مفاهیم سنتی نظیر جدول‌ها، سطر و ستون‌های ثابت بیگانه هستند و در بیشتر موارد، عملیات الحاق[3] در آن‌ها بی‌معنی بوده و به صورت افقی مقیاس‌پذیر هستند. با این حال، تا چند سال گذشته که تنها راه‌حل ذخیره‌سازی داده‌ها پایگاه داده رابطه‌ای بود، سیستم‌های نرم‌افزاری و پایگاه داده‌ای سنتی معماری مناسبی برای بهره‌برداری مناسب از مجموعه‌ای از ماشین‌ها را نداشته و محدودیت‌های ساختاری، آن‌ها را در مواجهه با ابعاد مختلف کلان داده‌ها شامل حجم، سرعت و تنوع، ناتوان و ناکارآمد ساخته بود. بر این اساس، چند سالی است که راه‌حل‌های مناسبی برای مدیریت و تحلیل این نوع داده‌ها تحت عنوان NoSQL مطرح شده است. امروزه تعداد زیادی از آنها به بلوغ رسیده و پتانسیل بسیاری برای استفاده و توسعه در محیط‌های کلان داده‌ها را دارند[۲, ۳].

بخش‌های مختلف مقاله به شرح زیر ارائه شده است. در ادامه، بعد از بررسی تحقیقات مرتبط در زمینه انتخاب پایگاه داده، در بخش دوم فرآیند اجرای پژوهش ارائه می‌شود. همچنین در این بخش به معرفی ساختار پایگاه‌های داده رابطه‌ای خواهیم پرداخت. در بخش سوم پایگاه‌های داده غیرساختاریافته مورد بررسی قرار می‌گیرند. در بخش چهارم ساختار فنی بانک‌های اطلاعاتی کلان داده‌ها مورد تحقیق قرار خواهد گرفت و در بخش پنجم ویژگی‌های کلیدی در انتخاب پایگاه‌های داده‌ی کلان داده معرفی خواهند شد. در بخش ششم، ساختار پایگاه داده‌ی ابری و راهکار مهاجرت به آن و در بخش هفتم شاخص‌های کارایی در پایگاه داده سنتی و توزیع شده ارائه می‌شود. در ادامه در بخش هشتم، مقایسه و انتخاب انواع پایگاه‌های داده با روش کارت امتیازی متوازن و در بخش نهم مقایسه کارایی ترکیب پایگاه داده سنتی و کلان داده مورد بررسی قرار می‌گیرد. در پایان نتیجه‌گیری و راهکارهای پیشنهادی ارائه شده است.

۱-۱- تحقیقات مرتبط

در این بخش تحقیقات مرتبط در زمینه انتخاب پایگاه‌های داده مناسب جهت ذخیره و بازیابی اطلاعات مورد بررسی قرار گرفته است. این تحقیقات برای کاربردهای مختلف بر اساس ساختار داده‌ای و نوع پایگاه داده صورت گرفته که به شرح زیر می‌باشد:

۱. در مرجع [۴]، در خصوص داده‌های با کاربرد بیومتریک، مقایسه و انتخاب یک پایگاه داده‌ی کلان مورد بررسی قرار گرفته است.
۲. در مرجع [۵]، به بررسی موضوع مقیاس‌پذیری در انتخاب پایگاه‌های داده SQL و NoSQL پرداخته شده است.
۳. در مرجع [۶]، به ارائه انتخاب یک پایگاه داده بر اساس مدل داده‌ای پرداخته شده است.
۴. در مرجع [۷]، به بررسی کنترل دسترسی و واسطهای دستیابی در پایگاه داده پرداخته شده است.
۵. در مرجع [۸] به بررسی تغییر الگو در پایگاه‌های داده ابری پرداخته شده است. این پایگاه داده یک بانک اطلاعاتی مبتنی بر سند بوده و در این تحقیق موضوع سرعت و مقیاس‌پذیری مورد بررسی قرار گرفته است.
۶. در مرجع [۹] به بررسی رویکرد مدیریت داده در مقیاس بزرگ در محیط‌های ابری پرداخته است. این نوع پایگاه‌های داده اشتراکی می‌باشند و پایگاه داده موجود در بستر ابر قرار می‌گیرد.
۷. در مرجع [۱۰]، به موضوع تصمیم‌گیری جهت انتقال و مهاجرت داده‌ها به لایه مرتبط در ابر پرداخته شده است. در اینجا مراحل انتقال در لایه پایگاه داده از سنتی به کلان داده مورد بررسی قرار گرفته است.
۸. در مرجع [۱۱]، برنامه‌های کاربردی در مدل‌های مختلف توسعه ابری، با استفاده از یک الگوی معماری چهار لایه‌ای ارائه شده است. این مدل‌ها شامل لایه نمایش، لایه کسب و کار، لایه دسترسی به داده و لایه پایگاه داده می‌باشد.

لازم به ذکر است که در هیچکدام از تحقیقات و کارهای ارائه شده مرتبط به موضوع انتخاب پایگاه‌های داده مناسب بر اساس حجم داده‌ها و ساختار داخلی آنها اشاره نگردیده است. همچنین ملاک در انتخاب پایگاه داده پارامترهای کیفی و کمی نبوده است، اما در این تحقیق ما با توجه به ویژگی‌های کیفی پایگاه‌های داده و بر اساس

[1] Carlo Strozzi
[2] Not Only Relational
[3] Join

ساختار داده‌های اطلاعاتی، موضوع انتخاب پایگاه داده مناسب را مورد بررسی قرار داده‌ایم.

۲- فرآیند اجرای پژوهش

در این تحقیق فرایند اجرای پژوهش در دو بخش اصلی مورد بررسی قرار گرفته است.

در بخش اول جنبه‌های نظری و تئوری موضوع مطرح شده است. از دید گاه نظری توسعه حجم و تنوع روز افزون داده‌ها در خدمات مختلف فناوری اطلاعات چالش‌ها و تگناهای جدی در تئوری‌های ذخیره‌سازی و تجهیزات ذخیره ساز سنتی ایجاد کرده است و بکارگیری این پایگاه‌های داده جهت ذخیره‌سازی داده‌های کلان را با چالش‌های مختلفی روبرو ساخته است. جهت حل این محدودیتها، ایجاد قابلیتهای تازه از جمله افزایش ظرفیت ذخیره‌سازی، قابلیت عملکرد موازی در فرایند پردازش و ذخیره‌سازی و ایجاد مدل‌های ذخیره‌سازی غیرساختاریافته پیشنهاد شده است. جهت اثبات میزان تاثیر راهکارارائه شده، ابتدا محدودیتهای پایگاه‌های داده رابطه‌ای و ویژگی‌های پایگاه‌های داده غیر رابطه‌ای تعیین شده است. به منظور دستیابی به راه حل مطلوب ساختار کلی بانک‌های اطلاعاتی جهت ذخیره‌سازی کلان داده‌ها پیشنهاد و مکانیزم مطلوب جهت انتخاب پایگاه داده مناسب برای کلان داده‌ها ارائه شده است.

در بخش دوم، به منظور اثبات راهکارهای ارائه شده، شاخص‌های کارایی در پایگاه داده سنتی و توزیع شده تعیین و کارایی ترکیبی پایگاه‌های داده سنتی و کلان داده با استفاده از روش کارت امتیازی متوازن مورد ارزیابی قرار گرفته و رفتار کلی سیستم و میزان کارایی پایگاه داده منتخب در حجم داده‌های مختلف محاسبه گردیده است. نتایج این بررسی محدودیت پایگاه‌های داده سنتی و چگونگی ارائه راهکار مناسب جهت ذخیره‌سازی داده‌های کلان در شبکه ملی اطلاعات را ارائه می‌نماید. در ادامه ما به تشریح بیشتر پایگاه‌های داده رابطه‌ای و غیررابطه‌ای پرداخته و ویژگی‌های آنها را بررسی خواهیم نمود.

۲-۱- ساختار پایگاه‌های داده رابطه‌ای

پایگاه داده سنتی مجموعه‌ای است از داده‌های ذخیره شده و پایا بر اساس یک مدل داده‌ای مشخص که به صورت مجتمع و یکپارچه به هم مرتبط بوده و با کمترین افزونگی تحت مدیریت یک سیستم متمرکز کنترل شده و توسط یک یا چند کاربر، به طور همزمان و اشتراکی مورد استفاده قرار می‌گیرد. معماری پایگاه‌های داده سنتی بصورت توزیع شده[1] و یا مرکزی[2] می‌باشد[۱۲]. ساختار این پایگاه‌های داده بصورت لایه‌ای، گرافی و یا جدولی می‌باشند. قوانین حاکم بر این پایگاه‌های داده دارای چهار ویژگی اصلی ACID[3] است. ACID متشکل از چهار ویژگی کلی به صورت زیر است که تضمین می‌کند تراکنش‌های پایگاه داده با قابلیت اطمینان خوبی پردازش می‌شوند[۳]:

- تجزیه‌ناپذیری (A): یک تراکنش باید به صورت کامل انجام شود و در غیر اینصورت باید لغو گرد.
- ثبات و سازگاری (C): یک تراکنش باید سیستم را از یک حالت سازگار به حالت سازگار دیگری ببرد و جامعیت داده‌ها را باتوجه به قوانین آن حفظ کند.
- جدا سازی (I): هر تراکنش باید از سایر تراکنش‌هایی که ممکن است همزمان در همان محیط در حال اجرا باشند مستقل باشد.
- ماندگاری (D): تاثیرات اجرای موفق یک تراکنش باید ماندگار باشند.

نمونه‌ای از این پایگاه‌های داده سنت عبارتند از MS SQL Server، IBM DB2، Oracle، MySQL و Microsoft Access. فراگیرترین مدل پایگاه‌های داده سنتی مدل رابطه‌ای می‌باشد. مدل‌های داده‌ای قبل از مدل رابطه‌ای در سطح پایینی از انتزاع قرار داشتند و وجود اشاره‌گرهایی که توسط طراح پایگاه داده سازماندهی می‌شدند کار طراحی را بسیار مشکل می‌کرد ولی با این وجود تنها راه دستیابی به کارایی مناسب در پردازش انبوه داده‌ها بودند. مدل رابطه‌ای در ابتدا به دلیل بار پردازشی بالا مورد استقبال سازندگان برنامه‌های مدیریت پایگاه‌های داده[4] قرار نگرفت ولی به تدریج و با قوی‌تر شدن کامپیوترها استقبال از آن بیشتر شد و امروزه تقریباً تمامی سیستم‌های مدیریت پایگاه‌های داده بر اساس آن ساخته می‌شوند. در این مدل زبان پرس‌وجوی استاندارد SQL مورد نظر قرار گرفته است. SQL، نامی است فراگیر برای رده‌ی گسترده‌ای از سامانه‌های مدیریت پایگاه‌های داده نوع سنتی. توجه به این نکته ضروری است که پایگاه‌های داده SQL همان مدل رابطه‌ای نیستند بلکه SQL یک زبان استاندارد رابطه‌ای است. SQL در بسیاری از موارد نمی‌تواند به خوبی از مدل رابطه‌ای پشتیبانی کند. این زبان از نظر ناقص بودن دستورات و نحوه اجرای آنها دارای مشکلات بسیار است.

[1] Distributed
[2] Central
[3] ACID (Atomicity, Consistency, Isolation, Durability)
[4] DBMS

تئوری زیر بنای مدل رابطه‌ای بر اساس یک اندیشه تئوریک به نام تئوری رابطه بنا شده است[۱]. یکی از موضوعات مهمی که سازندگان پایگاه داده در نظر دارند، دانش تئوری طراحی پایگاه داده است. مباحث نظریه‌ای رابطه توسط اکثریت سازندگان پایگاه داده پذیرفته شده است که بسیار مهم است. مدل رابطه‌ای وابسته به یک محصول خاص نیست بلکه وابسته به قوانین است. محصولات و فن‌آوری‌های رابطه‌ای مانندSQL و اوراکل مرتب در حال تغییر می‌باشند اما قوانین آنها ثابت هستند. بنابراین مدل رابطه‌ای و SQL یکی نیستند و تفاوت‌هایی با هم دارند. نکته دیگر این است که در مجموع مدل رابطه‌ای، طبیعتی اعلان‌گرا و نه روال‌گرا دارد. روش‌های اعلان‌گرا در پایگاه داده سنتی بر روش‌های روندگرا ترجیح دارند چرا که ملموس‌تر هستند[۱۳]. در حال حاضر، اصول ریاضی پشتیبان روش رابطه‌ای و مدل‌های پیاده‌سازی آن، ویژگی‌های خاصی دارد که مناسب داده‌های سنتی بوده و قابلیت حفظ و ذخیره‌سازی اطلاعات کم و متوسط را دارد. با گسترش ابزارهای جدید در صنعت فناوری اطلاعات و تغییر مدل و الگوهای آن، تناسب داده‌ها با روش‌های ذخیره‌سازی از تناسب خارج گردیده و سیستم‌های موجود کارایی خود را از دست داده‌اند.

۳- ساختار پایگاه‌های داده غیرساختار یافته

مفاهیم پایگاه‌های داده رابطه‌ای، مدت زمان زیادی بر برنامه‌های کاربردی مبتنی بر آنها حاکم و استفاده از آنها در برنامه‌های کاربردی بزرگ اجتناب‌ناپذیر کرده است. با این حال، برخی محدودیت‌ها و مشکلات باعث مطرح شدن مفاهیم جدیدی در پایگاه داده شده است. برای جواب دادن به نیازهای امروزی، پایگاه‌های داده‌ای جدیدی پا به عرصه وجود گذاشته‌اندکه با عنوان Not Only SQL یا NoSQL شناخته می‌شوند. محققان دانشگاهی این پایگاه‌های داده‌ای را با عنوان ذخیره‌سازی ساخت‌یافته می‌شناسند که مفهومی کلی بوده و در برگیرنده پایگاه‌های داده‌ای رابطه‌ای نیز هست. در سال ۲۰۰۹، شرکت Rackspace عبارت NoSQL را باز معرفی کرد و از آن برای اشاره به مجموعه‌ای از پایگاه‌های داده‌ای غیررابطه‌ای و گسترده در شبکه استفاده کرد که در نقطه مقابل پایگاه‌های داده‌ای رابطه‌ای قدیمی مانند Microsoft SQL Server، MySQL، IBM DB2، Informix، Oracle RDBMS، PostgreSQL، Oracle RDB و ... قرار می‌گرفتند[۱۲]. پس از بروز نیازهای جدید و معرفی تئوری‌های مورد نیاز برای پیاده‌سازی چنین پایگاه‌داده‌ای نوظهوری، محصولات مختلفی در این زمینه پا به عرصه وجود نهادند که هر یک از جهاتی مهم بوده و قابلیت بسیار سودمندی به همراه دارند. فعالیت‌های مرتبط با توسعه پایگاه‌های داده‌ای NoSQL در محافل علمی به شدت افزایش یافته است و اخیراً کار روی پروژه ای به نام UnQL[1] آغاز شده که هدف آن، تدوین استاندارد زبان پرس‌وجو در پایگاه‌داده‌ای NoSQL است[۲, ۳]. این زبان به گونه‌ای طراحی خواهد شد تا به جای جدول‌های سنتی و سطرها و ستون‌های داده‌ای، به ترتیب مجموعه‌ها[2]، اسناد[3] و فیلدهای مشخصی مورد پرس‌وجو قرار دهد. این زبان در اصل مجموعه‌ای فراتر از SQL به شمار آمده و زبان SQL را می‌توان نمونه بسیار محدود شده آن به شمار آورد.

۳-۱- کلیات پایگاه داده غیر رابطه‌ای NoSQL

اصطلاح NoSQL نامی عمومی است که به مجموعه‌ای از پایگاه‌های داده اطلاق می‌شود که از زبان پرس‌وجوی ساخت‌یافته SQL[4] یا مدل داده رابطه‌ای استفاده نمی‌کنند. این اصطلاح نخستین‌بار توسط اریک اوانس[5] به کار رفت. او که یکی از توسعه‌دهندگان پایگاه‌های داده غیررابطه‌ای است، بعدها او از به کار بردن این اصطلاح خودداری کرد و به جای آن اصطلاح کلان داده‌ها را به کار برد تا این گروه از پایگاه‌های داده را نه براساس عدم سازگاری با SQL، بلکه براساس مدیریت مقادیر کلان داده، تعریف کند. پایگاه‌های داده NoSQL، برخلاف مدل‌های رابطه‌ای یکسری ویژگی‌های خاصی ارائه می‌دهند و عبارتند از [۱۴, ۱۵]:

- نیازی به تعریف طرح[6] خاصی برای پایگاه داده نیست: داده‌ها می‌توانند در NoSQL بدون نیاز به تعریف الگوی محدود کننده‌ای ذخیره شوند. قالب داده‌هایی که در پایگاه داده NoSQL ذخیره می‌شوند بدون ایجاد آشفتگی در برنامه می‌توانند هر لحظه تغییر کنند. این ویژگی انعطاف‌پذیری بسیار بالایی به کاربردها می‌دهد که نیاز اساسی هر کسب‌وکاری است.
- بخش‌بندی خودکار[7]: این ویژگی با عنوان خاصیت کشسانی هم شناخته می‌شود. پایگاه داده NoSQL به صورت خودکار داده‌ها را بدون نیاز به دخالت برنامه بر روی سرورها منتقل

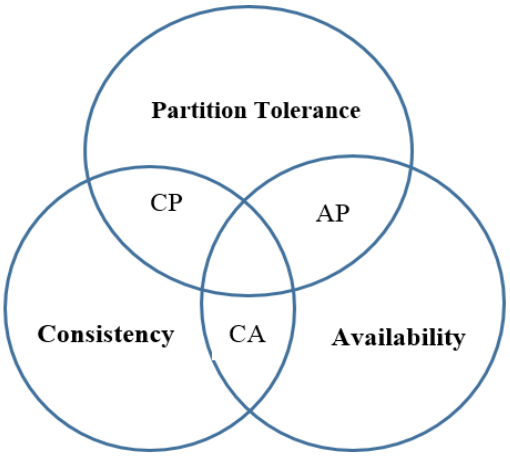
[1] Unstructured Query Language
[2] Collection
[3] Documents
[4] Structured Query Language
[5] Eric Evans
[6] Schema
[7] Auto-Sharding

می‌کند. در این حالت سرور می‌تواند بدون نیاز به برنامه لایه داده را مدیریت کند.

- پشتیبانی از پرس‌وجوی توزیعی: پایگاه داده NoSQL قدرت پرس‌وجو گرفتن از صدها یا هزاران سرور توزیع شده را دارد.
- حافظه نهان یکپارچه[1]: برای کاهش زمان تاخیر و افزایش بهره‌وری سیستم، تکنولوژی پیشرفته NoSQL قابلیت ذخیره کردن داده‌ها را در حافظه نهان سیستم دارد و این قابلیت به صورت شفاف انجام می‌شود.

۳-۲- مدل پایگاه‌های داده غیررابطه‌ای بر اساس تئوری CAP و ویژگیهای BASE

عبارت NoSQL یک مفهوم برای مشخص‌سازی یک تحول جدید است که در دنیای پایگاه‌های داده‌ای در حال وقوع است. هم‌اکنون با توسعه فناوری‌های مختلف و قابلیت نمونه‌برداری و تولید حجم عظیمی از داده‌ها، امکان ذخیره‌سازی و تحلیل آن‌ها چالشی بزرگ به شمار می‌آید. مدیریت داده‌هایی مانند داده‌های هواشناسی، فعالیت‌های برخط کاربران یا تحلیل‌های اقتصادی در قالب پایگاه‌های داده‌ای سنتی، کارایی چندانی نخواهند داشت. همچنین، امروزه سرویس‌دهندگان بسیاری به ذخیره‌سازی و ارائه محتوای کلان داده به کاربران خود در شبکه نیاز دارند که خود چالشی بسیار بزرگ به شمار می‌آید. کارآیی بسیار بالا در ذخیره‌سازی و ارائه داده‌هایی مانند اسناد PDF و فایل‌های MP3، در مقیاس وسیع، یکی از بهترین کاربردهایی است که پایگاه‌های داده‌ای NoSQL شایستگی خود را در فراهم‌کردن آن به اثبات رسانده‌اند. یک نمونه مناسب در این زمینه، خدمات Amazon S3 است. با این اوصاف، موارد ذکر شده تنها چالش پیش روی توسعه‌دهندگان و سرویس‌دهندگان نیست. ذخیره‌سازی، مدیریت و بازیابی داده‌های گذرا که در بعضی موارد در مقیاس بالایی در برنامه‌های کاربردی امروزی تولید می‌شوند نیز یکی دیگر از چالش‌های امروزی است که راه حل مدیریت مناسب آن‌ها را پایگاه‌های داده‌ای NoSQL ارائه‌کرده‌اند. این پایگاه‌های داده، در مدیریت داده‌هایی نظیر متغیرهای یک نشست[2] در وب، یک دیاگرام مناسب برای انتخاب راه حل درست توسط ناتان هورست[3] براساس نظریه CAP طراحی شده که در شکل ۱ آمده است[۲, ۱۶].



CP	پایگاه های داده با خاصیت بخش‌بندی و پایداری و ACID مانند گرافی، ستونی و HBASE
AP	پایگاه های داده با خاصیت قابلیت دسترسی و پایداری و BASE مانند کلید/ مقدار، نوع مسند، کاساندرا.
CA	پایگاه های داده با خاصیت دردسترس پذیری و پایداری مانند اوراکل و SQL.

شکل ۱. دیاگرام نظریه CAP

در دیاگرام شکل ۱ سه گوشه اصلی نشانگر ویژگی‌های ثبات[4]، دردسترس‌بودن[5] و قابلیت بخش‌بندی[6] هستند. ثبات در اینجا یعنی همه کاربران همواره به داده‌های مشابه دسترسی داشته باشند، دردسترس بودن یعنی همه کاربران امکان خواندن و نوشتن را داشته باشند و قابلیت بخش‌بندی نیز به معنای این است که سیستم به رغم تقسیم شدن فیزیکی شبکه به قسمت‌های مختلف به خوبی کار کند. بر اساس نظریه CAP، تنها دو عنصر از این سه عنصر می‌توانند همزمان در سیستم‌های واقعی به صورت قطعی و تضمینی وجود داشته باشند. بر همین اساس، برای داشتن هر جفت مشخصه، می‌توان راه‌حلی را که روی ضلع مشترک آن‌ها آورده شده است، انتخاب کرد.

همچنین ویژگی BASE[7] در طراحی پایگاه‌های داده غیررابطه‌ای و در مقایسه با سیستم‌های سازگار با ACID مطرح می‌شود. این اصطلاح برای توصیف ویژگی‌های انواع تراکنش‌هایی است که توسط پایگاه‌های داده غیررابطه‌ای پشتیبانی می‌شوند. یک سیستم BASE، با سه ویژگی مشخص تشریح می‌شود که عبارتند از: اساساً در دسترس بودن، وضعیت نرم و در نهایت پایدار بودن. در پایگاه‌داده BASE، گزینه اول که در دسترس بودن یا مهیا بودن است ضروری‌ترین خصیصه می‌باشد. یک نمونه از پایگاه‌های داده مبتنی

[1] Integrated caching
[2] Session
[3] Nathan Hurst
[4] Consistency
[5] Availability
[6] Partition Tolerance
[7] Basically Available, Soft state, Eventual consistency

بر BASE، پایگاه داده مربوط به شبکه‌های اجتماعی است که در مقایسه با پایگاه‌های داده بانکی که به طور سرسخت بر پایه تئوری CAP استوار هستند، حساسیت زیادی در وجود ویژگی ثبات ندارد. لذا در پایگاه داده شبکه اجتماعی مبتنی بر خصیصه BASE عمل می‌شود[۱۲].

۴- ساختار فنی بانک‌های اطلاعاتی داده‌های کلان

اگر چه NoSQL به عنوان پایگاه‌های داده‌ی کلان داده پذیرفته شده است ولی یک راه‌حل جهت حل تمام مشکلات نمی‌باشد. در این قسمت نمونه‌هایی از پایگاه‌های داده برای کلان داده‌ها متداول مورد بررسی قرار خواهند گرفت و به ویژگی‌های فنی آنها اشاره خواهد شد.

- Amazon SimpleDB: این پایگاه داده غیررابطه‌ای بسیار دسترس‌پذیر، مقیاس‌پذیر و انعطاف‌پذیر می‌باشد. یک واسط کاربری ساده با Get، Post، Delete و اجرای پرس‌وجو را ارائه می‌دهد. همچنین به کاربر اجازه می‌دهد قبل از ذخیره‌سازی داده‌ها آنها را رمزگذاری نمایند
- Google App's Bigtable: این پایگاه داده یک سیستم ذخیره‌سازی توزیع شده بر مبنای GFS برای داده‌های ساخت‌یافته می‌باشد. علاوه بر آن در بسیاری از محصولات گوگل مانند Google app engine به خوبی پیاده‌سازی شده است و اجازه می‌دهد تا داده‌های پیچیده ذخیره شوند.
- Hadoop: یک زیرساخت برنامه‌نویسی برای پیاده‌سازی مدل MapReduce بر روی سیستم‌های توزیع شده می‌باشد و بیشتر مناسب داده‌های بدون ساختار می‌باشد.
- CouchDB: یک محصول متن‌باز در پروژه Apache است که از سال ۲۰۰۸ معرفی شده است. در این پایگاه داده که یک بانک اطلاعاتی سندگرا می‌باشد، رکوردها با فرمت JSON[1] ذخیره می‌شوند و با استفاده از رابط HTTP بازیابی می‌گردند.
- MongoDB: یک بانک اطلاعاتی سندگرا است و به عنوان یک بانک اطلاعاتی شیءگرا طراحی شده است. این پایگاه داده دارای سرعت و مقیاس‌پذیری بالایی می‌باشد[۱۷, ۱۸].

۵- ویژگی‌های انتخاب پایگاه کلان داده‌ها

در حالت کلی مقایسه و انتخاب یک پایگاه داده‌ی کلان داده‌ها بر اساس پنج مشخصه اصلی زیر می‌باشد[۴]:

- مقیاس‌پذیری [۵]
- پایداری و یکپارچگی در تراکنش‌ها [۱۹]
- مدل داده‌ای [۶]
- پشتیبانی از پرس وجو[2] [۲۰]
- کنترل دسترسی و واسطهای دستیابی [۷]

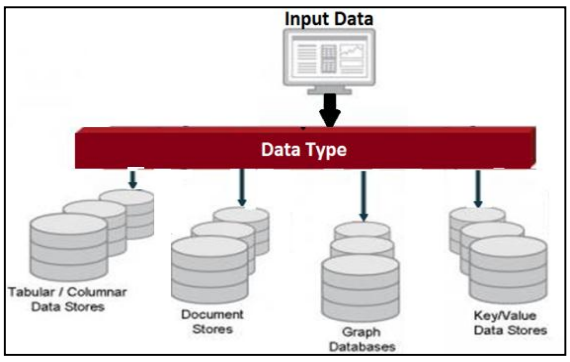
اما با توجه به این شاخص‌ها، محصولات نرم‌افزاری زیادی برای انتخاب کردن وجود دارند که می‌بایست بر اساس دسته‌بندی‌های مشخصی آنها را شناسایی کرد. می‌توان به ترتیب از روش‌های زیر برای این کار استفاده کرد که در ادامه به آن می‌پردازیم.

۵-۱- انتخاب بر اساس نوع داده و تناسب آن با ویژگی‌های پایگاه‌داده

در این روش پایگاه داده با توجه به نوع داده‌های اطلاعاتی و ویژگی‌های آنها انتخاب می‌شود. شکل ۲ فرآیند انتخاب پایگاه داده بر اساس تطابق ویژگی‌های کلیدی داده و مشخصات پایگاه داده را نشان می‌دهد.

در این روش انتخاب پایگاه داده می‌تواند به یکی از چهار حالت زیر صورت پذیرد:

نوع اول: داده‌هایی که می‌توانند با مدل کلید-مقدار[3] اندیس‌گذاری شوند: این روش شبیه نقشه‌ها یا دیکشنری‌ها هستند که درون آنها داده‌ها با یک کلید منحصر بفرد شناسایی می‌شوند. نمونه‌هایی از پایگاه‌های داده از این گروه عبارتند از: Redis، Riak و MemcacheDB.



شکل ۲. انتخاب پایگاه داده مناسب بر اساس نوع داده

[1] JavaScript Object Notation

[2] Query

[3] Document-oriented

نوع دوم: داده‌هایی که می‌توانند با مدل مبتنی بر سند ذخیره شوند: در این مدل، اطلاعات بصورت جفت‌های کلید و مقدار در فرمت پایگاه‌داده غیررابطه‌ای JSON یا شبیه آن ذخیره می‌شود. درون اسناد کلیدها باید منحصربه‌فرد باشند. در مقایسه با مدل کلید-مقدار، در این مدل مقادیر برای سیستم قابل فهم هستند و می‌توان بر روی آنها پرس‌وجو اجرا کرد. نمونه‌هایی از پایگاه‌های داده از این گروه MongoDB و Couchbase می‌باشند.

نوع سوم: داده‌هایی که می‌توانند با مدل مبتنی بر ستون[1] ذخیره شوند: این مدل، از ستون‌ها برای ذخیره‌سازی اطلاعات استفاده می‌کند. نمونه‌هایی از پایگاه‌های داده از این گروه Cassandra و HBase می‌باشند.

نوع چهارم: داده‌هایی که می‌توانند با مدل مبتنی بر گراف ذخیره شوند: این مدل به‌صورت تخصصی برای مدیریت داده‌هایی که به یکدیگر متصل هستند بهینه شده است. نمونه‌هایی از پایگاه‌های داده این گروه Neo4J و OrientDB می‌باشند[۲, ۲۱].

۵-۲- انتخاب بر اساس ویژگی‌های فنی پایگاه داده

در این روش پایگاه‌های‌داده بر اساس تئوری CAP انتخاب می‌گردند. در این ساختار سه ویژگی اصلی C، A وP سه رأس تصمیم‌گیری هستند و بر اساس نظریه CAP، تنها دو ویژگی از این سه ویژگی در سیستم‌های واقعی قابل تضمین هستند. شکل ۱ ترکیب ویژگی‌ها و پایگاه‌های داده سازگار متناسب با نوع داده را نشان می‌دهد. به عنوان مثال اگر ویژگیهای C و A مورد نظر باشند می‌توان یکی از پایگاه‌های داده نوع RDBMS را انتخاب کرد. اگر ویژگی‌های C و P مورد نظر باشند می‌توان یکی از پایگاه‌های داده MongoDB یا HBaseرا انتخاب کرد. همچنین اگر ویژگیهای P و A مورد نظر باشند می‌توان یکی از پایگاه‌های داده Cassndra یا CouchDBرا انتخاب کرد[۱۶].

۵-۳- انتخاب بر اساس ویژگیهای سرویس ابر

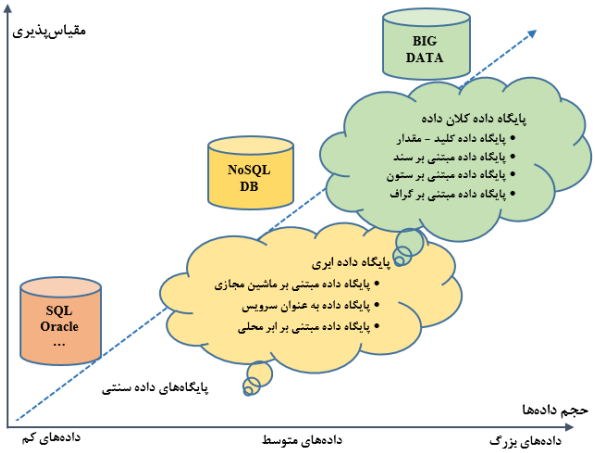
پایگاه‌های داده غیررابطه‌ای می‌توانند به صورت پیش‌فرض در محیط ابر اجرا شوند اما اغلب بر روی بسترهای IaaS و PaaS اجرا می‌شوند. در این روش بر اساس ساختار و نحوه ارائه سرویس می‌توان مدل پایگاه داده ابری را انتخاب نمود. سه انتخاب برای یک پایگاه داده ابری وجود دارد که عبارتند از:

- پایگاه داده‌ی مبتنی بر ماشین مجازی: پایگاه‌داده به عنوان بخشی از مجازی‌سازی مستقر می‌شود و چندین ماشین مجازی، سرور فیزیکی یکسانی را به اشتراک می‌گذارند. این روش ماکزیمم سطح جداسازی (در سطح سیستم عامل) را ارائه می‌کند و در نتیجه رشد بی‌رویه ماشین‌های مجازی یکسری مسائل امنیتی در نظر می‌باشد.
- پایگاه داده به عنوان سرویس ابری: پایگاه‌داده به عنوان یک سرویس، منابع را در اختیار کاربر یا کاربران قرار می‌دهد. این روش به کاربران اجازه می‌دهد یک یا چند پایگاه‌داده را تعریف و مدیریت کنند. هر مشترک می‌تواند فقط پایگاه‌داده‌ای را که به او اختصاص داده شده است ببیند و پایگاه‌های داده مشترکان دیگر برای او قابل روئیت نیست. همچنین مدیریت و هماهنگی‌های پایگاه داده را خود بر عهده دارد و می‌تواند منابع محاسبه اضافی و ذخیره‌سازی مبتنی بر تقاضا را از طریق پرتال ابری انتخاب و اختصاص دهد. از جمله این پایگاه‌های داده MongoDB و Cassandra می‌باشند[۲۲].
- پایگاه داده به عنوان سرویس پایه ابر[2]: یکسری از فراهم‌کنندگان ابر، این سرویس را به صورت پیش فرض در دسترس قرار می‌دهند. این پایگاه داده به عنوان سرویس پایه خدمات ابر نامیده می‌شود. همچنین این پایگاه داده اشتراکی است و خدمات پایگاه داده بر روی طرح پایگاه داده موجود استقرار می‌یابند. Amazon SimpleDB یکی از مشهورترین این نوع پایگاه داده است[۹, ۲۳].

شکل ۳ انتخاب پایگاه‌های داده‌ی مناسب بر اساس ساختار و حجم داده را بطور مرحله به مرحله بر اساس سه مدل انتخابی نشان می‌دهد. با توجه به تنوع ویژگی داده‌ها، متدولوژی انتخاب پایگاه داده مناسب به روش‌های مختلفی میسر می‌باشد، در اینجا برای داده‌های با حجم کم انتخاب پایگاه‌های داده سنتی مناسب می‌باشد و در استفاده از محیط‌های مجازی و خدمات ابری پایگاه‌های داده NoSQL راهکار مناسبی را ارائه می‌نماید. در داده‌های کلان با حجم زیاد انتخاب پایگاه‌های داده مناسب تنوع مختلفی داشته و بر اساس نوع داده گزینه‌های مختلفی از مدل‌های سنتی تا مدل‌های مبتنی بر کلید مقدار، ستونی، گرافی و مبتنی بر سند قابل انتخاب می‌باشند که از روی ویژگی‌های داده به طور خودکار انتخاب خواهند گردید[۲۹].

[1] Column-oriented

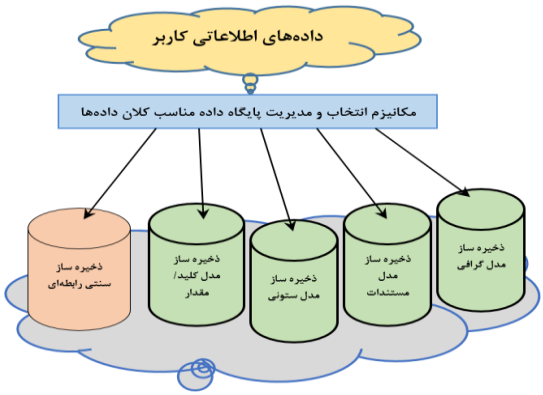
[2] Native



شکل ۳. متدولوژی انتخاب پایگاه داده مناسب بر اساس حجم داده

۵-۴- مکانیزم پیشنهادی جهت انتخاب پایگاه داده مناسب برای کلان داده‌ها

چالش اصلی که دنیای کلان داده‌ها با آن مواجه می‌باشند، این است که یک حجم مشخص داده می‌تواند شامل اجزای مختلفی باشد که هر یک باید در یکی از انواع پایگاه‌های داده ذخیره شود و هیچ پایگاه داده واحدی برای ذخیره‌سازی همه آنها وجود ندارد. مدل ارائه شده بر روی این مسئله کار می‌کند که چگونه داده‌های ورودی بین پایگاه‌های داده مختلف تقسیم شود به نحوی که امکان پردازش مناسب آن‌ها میسر باشد. در این بخش به ارائه یک مکانیزم پیشنهادی جهت انتخاب پایگاه داده مناسب برای کلان داده‌ها می‌پردازیم. در اینجا فرض بر وجود یک حجم داده اطلاعاتی با ترکیبات مختلف می‌باشد و هدف انتخاب پایگاه داده مناسب جهت ذخیره‌سازی آن اجزاء است. انتخاب‌های ممکن و نحوه انتخاب پایگاه داده‌ی کلان داده‌ها مناسب در شکل ۴ نشان داده است.



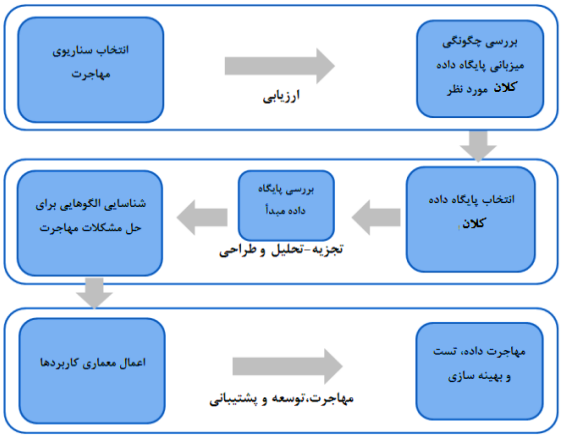
شکل ۴. مدل ترکیبی[1] در انتخاب پایگاه کلان داده مناسب

این انتخاب با توجه به مشخصات فنی پایگاه‌های داده که در بخش‌های قبل آمده است، انجام گرفته است. مکانیزم انتخاب و مدیریت پایگاه‌های داده در نمونه پیشنهادی بر اساس نتایج ارزیابی صورت گرفته بین پایگاه‌های داده سنتی و کلان داده‌ها و همچنین متریک‌های فنی انواع مختلف پایگاه‌های داده‌ای کلان داده‌ها است. در الگوی انتخاب، پنج مدل پایگاه داده با ویژگی‌های متفاوت شامل پایگاه داده رابطه‌ای یا سنتی، پایگاه‌های داده بر اساس کلید-مقدار، پایگاه‌های داده ستونی، پایگاه داده محتوایی و پایگاه داه گرافی می‌باشند. متناسب با ویژگی‌های داده‌های ورودی و مکانیزم انتخاب، پایگاه داده مناسب برای داده مورد نظر انتخاب می‌گردد. این انتخاب مادامی که ویژگی‌های داده‌ها تغییر نکند ثابت بوده و با تغییر پارامترهای کلیدی داده ممکن است تغییر نماید[۱۲, ۲۴].

۶- ساختار پایگاه داده‌ی ابری و راهکار مهاجرت به آن

به طور کلی، انتقال لایه داده برنامه‌های کاربردی به پایگاه داده‌ی کلان داده‌ها، یک مسئله پیچیده و چندبعدی است. جهت انتقال لایه داده به بانک کلان داده، چالش‌هایی ایجاد می‌گردد که جهت رفع این چالش‌ها از روش مرحله به مرحله[2] استفاده می‌شود.

در اینجا ما روش لازوسکی و نادوری[3]، که شامل هفت مرحله جداگانه می‌باشد را مطابق شکل ۵، ارائه می‌نماییم[۱۰]. این روش برای انتقال لایه پایگاه داده از سنتی به پایگاه داده‌ی کلان داده‌ها است که مراحل آن مشخص شده است. این مراحل به شرح زیر می‌باشد:



شکل ۵. مراحل انتقال در لایه پایگاه داده از سنتی به کلان داده‌ای [۱۰]

[1] Hybrid

[2] Step-by-step

[3] Laszewski, Nauduri

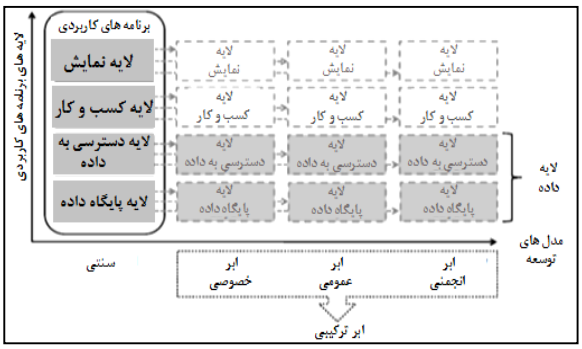
بخش ارزیابی، اطلاعات مرتبط با مدیریت پروژه مانند ابزارها و گزینه‌های مهاجرت به منظور ارزیابی تأثیر مهاجرت بانک اطلاعاتی جمع‌آوری می‌گردد.

بخش تجزیه و تحلیل جزئیات پیاده‌سازی در بانک اطلاعاتی هدف، یعنی انواع داده‌های مختلف، مکانیزم‌های مدیریت تراکنش را بررسی می‌کند. در این بخش انتخاب پایگاه داده مناسب انجام می‌شود و یک طرح را برای اجرای آن می‌سازد.

بخش مهاجرت این بخش به مهاجرت داده‌ها از پایگاه داده مبدأ به پایگاه داده مقصد در محیط تستی می‌پردازد. بعد از مرحله مهاجرت در مرحله تست هم بانک اطلاعاتی و هم نرم‌افزار بررسی می‌گردند که شامل تایید صحت داده نیز می‌باشد. هر گونه بهینه‌سازی بر اساس پایگاه‌داده مقصد اعمال می‌شود. در نهایت در مرحله استقرار سیستم نهایی شامل بانک اطلاعاتی مهاجرت یافته در محیط واقعی مستقر می‌گردد.

جهت بهره‌گیری از مزایای تکنولوژی رایانش ابری، بکارگیری از پایگاه‌های داده مبتنی بر رایانش ابری برای کلان داده‌های به صورت فزآینده‌ای مورد توجه قرار گرفته است. برای استفاده از مزیت‌های رایانش ابری، ممکن است برنامه‌های کاربردی موجود به رایانش ابری منتقل شوند یا از همان ابتدا به صورت ابری طراحی گردند. برنامه‌های کاربردی مطابق شکل ۶، در مدل‌های مختلف توسعه ابری، با استفاده از یک الگوی معماری چهار لایه‌ای ساخته می‌شوند[۱۱]:

۱. **لایه نمایش**، که تعاملات بین کاربر و کاربرد را تشریح می‌کند.
۲. **لایه کسب‌وکار**، که منطق کسب‌وکار را تحقق می‌بخشد.
۳. **لایه دسترسی به داده** (DAL[1])، که مسئول ذخیره‌سازی داده‌های کاربرد می‌باشد.



شکل ۶. ساختار لایه‌های نرم‌افزاری در پایگاه داده ابری

۴. **لایه پایگاه‌داده** (DBL[2])، که مسئول تدوام و نگهداری داده‌ها می‌باشد.

در اینجا مهاجرت لایه داده سنتی به ابری شامل دو گام اساسی می‌باشد:

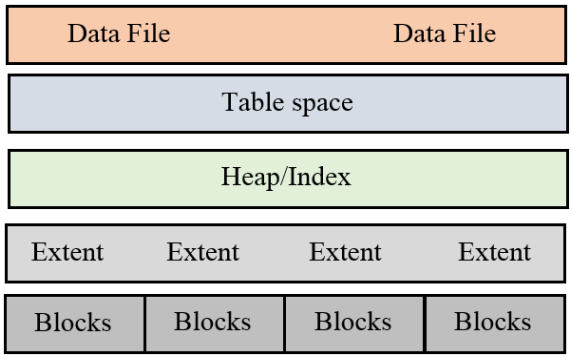
- مهاجرت DBL به ابر
- سازگاری DAL جهت فعال‌سازی دسترسی

همچنین شکل ۶، موقعیت لایه‌های مختلف برای توزیع یک کاربرد در مدل‌های مختلف توسعه ابری را نمایش می‌دهد. لایه داده مجموع دو لایه دسترسی به داده و لایه پایگاه‌داده می‌باشد. DAL مسئول دسترسی به داده‌ها می‌باشد در حالی که DBL مسئول تدوام و نگهداری داده‌ها می‌باشد [۳۰].

۷- شاخص‌های کارایی پایگاه‌های داده سنتی و توزیع شده

در بخش ششم، روش‌های مهاجرت به پایگاه داده‌های کلان مورد بحث قرار گرفت. در اینجا به بررسی شاخص‌های کارایی در پایگاه‌های داده سنتی و توزیع شده خواهیم پرداخت. جهت این بررسی، عملکرد پایگاه‌های داده سنتی و توزیع شده هدوپ بر اساس ویژگی‌های آنها مورد مقایسه قرار گرفته است. شکل ۷ ساختار داخلی پایگاه‌های داده سنتی را نشان می‌دهد. در این پایگاه‌های داده اصول ذخیره‌سازی بر اساس جداول طراحی شده صورت می‌گیرد.

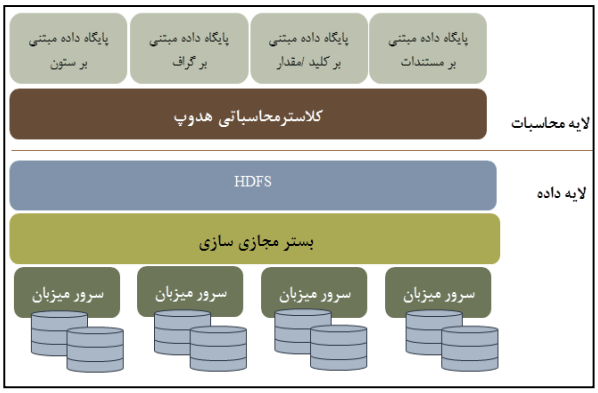
در شکل ۸ ساختار لایه‌بندی پایگاه داده توزیع شده هدوپ و نحوه ارتباط زیر بخش‌های مختلف لایه داده و لایه محاسبات آن نشان داده شده است.



شکل ۷. ساختار پایگاه داده سنتی رابطه‌ای

[1] Data Access Layer

[2] Database Layer



شکل ۸. ساختار پایگاه داده‌ی کلان داده‌ها در بستر هدوپ

بر اساس این شکل، انتخاب نوع پایگاه داده می‌تواند بر اساس مشخصات فنی و مقایسه مراحل اجرایی هر یک از پایگاه‌های داده سنتی و توزیع شده که در اینجا تشریح شده‌اند، انجام گیرد. به منظور مقایسه عملکرد پایگاه‌های داده مختلف ویژگی‌های مختلف آنها مطابق جدول شماره ۱ مورد بررسی قرار گرفته است.

جدول ۱. مقایسه ویژگی‌های پایگاه‌های داده سنتی و کلان داده

	سنتی	کلان داده‌ها
مدل محاسباتی	- بر اساس مفهوم تراکنش - تراکنش‌ها واحد های کاری هستند - مبتنی بر ACID	- بر اساس مفهوم Jobها - Jobها واحدهای کاری‌اند - مبتنی بر CAP و BASE
مدل داده‌ای	- داده‌های ساخت یافته با طرحی مشخص	- داده‌ها با فرمت دلخواه - داده‌های بدون ساختار/ نیمه ساختار
مدل توزیع شدگی	- استفاده از سرورهای متمرکز	- کامپیوترها و سرورهای توزیع شده
تحمل خطا	- خطا به ندرت ایجاد می‌شود - دارای مکانیزم‌های بهبود و ترمیم	- خطاها بر روی ماشین‌های زیادی ایجاد می‌شوند. - تحمل خطای کارآمد
میزان تاخیر	با تاخیر کم در حجم داده کم	با تاخیر زیاد در حجم داده زیاد
مقیاس‌پذیری	مقیاس پذیری عمودی	مقیاس پذیری افقی
مشخصات کلیدی	- کارایی، بهینه سازی، تنظیم خوب	- مقیاس‌پذیری، انعطاف‌پذیری، تحمل‌خطا

این جدول، مقایسه کلی از ویژگی‌های کلیدی پایگاه‌های داده سنتی و کلان داده‌ها را نشان می‌دهد. فاکتورهای مقایسه شده بر اساس پارامترهای کلیدی دخیل در انتخاب و مدیریت پایگاه‌های داده انتخاب شده‌اند و نتایج ارزیابی‌های صورت گرفته بین پایگاه‌های داده سنتی و کلان داده‌ها ارائه شده است.

از مسائل مهم دیگر در انتخاب پایگاه‌های داده، مسئله مقیاس‌پذیری در ابعاد افقی و عمودی می‌باشند که در طراحی و اجرای سامانه‌های پیچیده و بزرگ همچون پایگاه‌های داده بسیار حیاتی و محوری می‌باشد. مقیاس‌پذیری را می‌توان در ابعاد گوناگونی اندازه‌گیری کرد. مقیاس‌پذیری عمودی به معنی گسترش منابع در یک گره از سیستم ذخیره‌سازی، به طور نمونه ارتقای پردازنده یا سامانه ذخیره‌سازی در یک کامپیوتر می‌باشد. از طرف دیگر، مقیاس‌پذیری افقی به معنی افزایش تعداد گره‌ها و تجهیزات ذخیره‌سازی بیشتر به سیستم است. میزان تاثیر انواع مقیاس‌پذیری افقی بیشتر از عمودی می‌باشد. مطابق قانون امدال[1] با دو برابر شدن قدرت پردازش، سرعت اجرا تنها به میزان یک پنجم افزایش می‌یابد. این بدین معنی است که افزودن سخت‌افزار در یک سیستم لزوماً شیوه بهینه‌ای جهت افزایش کارایی نیست. البته اگر کل برنامه قابل موازی‌سازی بود انتظار می‌رفت که سرعت نیز دو برابر گردد.

۸- مقایسه انواع پایگاه‌های داده با روش کارت امتیازی متوازن

یکی از روش‌های متداول در انتخاب و ارزیابی روش‌های مختلف در خدمات فناوری اطلاعات، استفاده از روش کارت امتیازی متوازن[2] می‌باشد. کارت امتیازی متوازن یا BSC، روشی برای مدیریت عملکرد یک سیستم است؛ که ایده اولیه آن سال ۱۹۹۲، در تحقیقات رابرت کاپلان و دیوید نورتون[3]، در زمینه روش‌های نوین سنجش عملکرد سازمان‌ها شکل گرفت. کارت امتیازی متوازن، یک ابزار مدیریتی برای اجرای استراتژی است؛ گزارش ساختاربندی شده و استانداردی که به مدیر یا سیستم اجازه می‌دهد بتوانند به راحتی بر روند اجرای فعالیت‌ها نظارت داشته باشند و نتایج این فعالیت‌ها را بررسی و کنترل کنند. ویژگی اصلی کارت امتیازی متوازن در اختیار گذاردن بستری مناسب برای شناخت قوانین و روابط علت و معلولی حاکم بر دنیای کسب و کار و همچنین استخراج برنامه‌های عملیاتی برای اجرایی کردن استراتژی در یک سیستم یا سازمان است[۲۵, ۲۶]. برنامه‌ریزی راهبردی شامل بررسی قابلیت‌ها و ضعف‌های محیطی و درونی می‌شود به طوری که نتایج حاصل از انتخاب استراتژی را به عملیات‌های روزمره پیوند می‌زند. کارت امتیازی متوازن با چنین هدفی توسعه یافته است که شاخص‌ها و متریک‌ها در برنامه‌ریزی راهبردی تدوین شوند و آنها عملیاتی و قابل اندازه‌گیری باشند. در این پژوهش نیز به‌منظور مدیریت و انتخاب دقیق پایگاه‌های داده مناسب، از روش کارت امتیازی متوازن با در نظر گرفتن پارامترهای مختلف از منظر داشتن نوآوری، قابلیت اجرا در فرآیندهای داخلی، رضایت کاربر و مباحث هزینه‌ای استفاده شده

[1] Amdahl's law

[2] Balanced Score Card (BSC)

[3] Robert S.Kaplan and David P.Norton

است. این پارامترها برگرفته از ویژگی‌های پایگاه‌های داده مختلف بوده و نقش مهمی در توانایی‌های آنها ایجاد می‌کند. جهت انتخاب مناسب پوشش حداکثری پارامترهای کلیدی می‌تواند نقش تعیین کننده‌ای را ایفا نماید. بدین منظور با توجه به نیازهای اساسی بر پایه ویژگی‌های داده‌ها، پارامترهای مختلف کلیدی در این زمینه مشخص و با توجه به درجه اهمیت آنها امتیاز مشخصی در نظر گرفته شده که مبین ویژگی‌های کیفی آنها می‌باشند.

به منظور تبدیل ویژگی‌های کیفی به مقادیر کمی سطوح پنجگانه تعریف شده است که هریک درجه مشخصی از مقادیر کیفی را نشان می‌دهد. برای تفکیک سطوح مختلف سطح ضعیف، پایین، متوسط، بالا، عالی در نظر گرفته شده است. مقدار کیفی مطابق جدول ۲ امتیاز دهی شده است. همچنین ویژگی‌هایی که تاثیر مثبت دارند با مقایر مثبت و ویژگی‌هایی که تاثیر منفی دارند با مقایر منفی می‌باشند.

جمع‌بندی و انتخاب متریک‌های فنی موردنظر برای ارزیابی انوع مختلف پایگاه داده شامل مقیاس‌پذیری [۵]، انعطاف‌پذیری [۱۴, ۱۵]، پیچیدگی [۲۴]، سرعت اجرا [۵]، حجم داده‌ها [۲۱]، تنوع داده‌ای [۲۱]، تحمل خطا [۲۷]، دسترسی‌پذیری [۱۲]، ثبات [۳] و بخش‌بندی [۱۲] می‌شود. فرایند اجرای BSC، بر اساس متریک‌های فنی مرتبط با انواع مختلف پایگاه‌های داده ساختار یافته و غیر ساختار یافته و با توجه به مستندات و منابع مختلف فنی در این زمینه انجام شده است. همچنین تجربیات پروژه‌های انجام شده مرتبط با پایگاه داده به عنوان راهکارهای کلیدی مورد استفاده قرار گرفته است. امتیازدهی و نتایج ارزیابی‌های بدست آمده در جدول ۳ ارائه شده است. در این جدول، برای تعیین امتیاز نهایی از روش کارت امتیازی متوازن BSC با در نظر گرفتن وزن یکسان برای همه متریک‌ها، محاسبات صورت گرفته و مقادیر بدست آمده درج گردیده است. در استراتژی کارت امتیازی متوازن برای انتخاب پایگاه داده، همه جنبه‌های کلیدی از منظرهای مختلف نوآوری، فرآیندهای داخلی، کاربری و هزینه‌ای با پارامترهای استانداردی مهم در ارزیابی پایگاه‌های داده استفاده شده است.

جدول ۲. مقادیر کیفی و امتیازهای در نظر گرفته شده

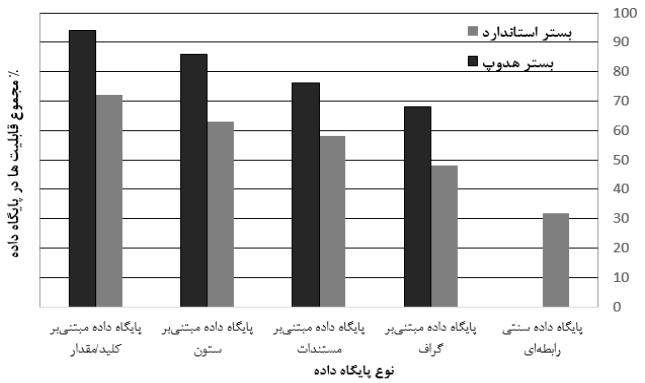
مقدار کیفی	امتیاز
ضعیف	±۱
عملکرد پایین	±۲
متوسط	±۳
بالا	±۴
عالی	±۵

بنابراین روش پیشنهادی به جای تمرکز تنها بر قسمتی از قابلیت یک پایگاه داده، یک دید کلی از عملکرد پارامترهای مختلف را برای انتخاب در نظر می‌گیرد. مطابق اطلاعات ارائه شده در جدول شماره ۳، دو بستر استاندارد و هدوپ انتخاب و پنج پایگاه داده مختلف با یازده متریک کلیدی متفاوت امتیازدهی گردیده که مجموع نتایج امتیازات ارائه شده است. بر این اساس، شکل ۹ درصد کارایی مجموع قابلیت‌های جدول ۳ برای انواع مختلف پایگاه‌های داده را به صورت نمودار میله‌ای نشان می‌دهد.

همان‌طورکه در نتایج این جدول مشخص شده است حضور زیرساختار هدوپ برای پایگاه داده، قابلیت‌ها در تمامی متریک‌های مختلف بهبود می‌یابند. هدوپ برای ذخیره‌سازی و پردازش توزیع شده مجموعه‌ی کلان داده‌ها از مدل برنامه‌ریزی کلید/مقدار استفاده می‌کند. این چارچوب به صورت خوشه‌هایی از سخت‌افزارها ایجاد می‌شود. همه ماژول‌ها در هدوپ با این فرض اساسی طراحی شده‌اند که در صورت خرابی بتواند به صورت خودکار مشکلات احتمالی را برطرف کنند تا مجموعه سیستم بتواند بصورت پیوسته و مستمر خدمات ارائه نماید. همچنین جدول ۳ مقایسه نتایج ارزیابی در متریک‌های متفاوت را نشان می‌دهد.

۹- مقایسه کارایی ترکیب پایگاه داده سنتی و کلان داده

به منظور تعیین رفتار پایگاه‌های داده متفاوت در مواجه با داده‌های ورودی، بایستی عملکرد آنها را برای داده‌های مختلف در نظر بگیریم. با توجه به تجربیات و مطالب ارائه شده در بالا، قابلیتهای کارایی در پایگاه‌های داده سنتی بیشتر در سرعت اجرا، دسترسی‌پذیری، ترمیم و همزمانی تمرکز داشته و این کارایی نسبت عکس با میزان مربع حجم داده‌های ورودی دارد. بطوری‌که چنانچه حجم داده‌های ورودی افزایش یابد، کارایی پایگاه داده به‌شدت کاهش می‌یابد.



شکل ۹. مقایسه کارایی پایگاه‌های داده منتخب جهت ذخیره‌سازی کلان داده‌ها در محیط‌های استاندارد و بستر هدوپ

بنابراین با توجه به نمودارها مراجع [۷] [۲۸] تناسبی که می‌توان میان کارایی یک پایگاه داده سنتی و حجم داده ورودی در نظر گرفت به صورت P_{tr} در رابطه ریاضی (۱) خواهد بود:

$$P_{Tr} = K(1 - v^2), \quad (۱)$$

در این رابطه v حجم داده و K مقدار ثابت می‌باشد. از طرف دیگر با توجه به مطالب بیان شده قبل، کارایی پایگاه داده‌های توزیع شده برای کلان داده‌ها شامل ارزانی، مقیاس‌پذیری، تحمل‌پذیری، انعطاف، بخش‌پذیری، تنوع داده‌ای و حجم داده می‌باشند که در مجموع ۷ قابلیت را شامل می‌شود. در عمل کارایی پایگاه داده‌ی کلان داده‌ها برای داده‌های گسترده بیشتر از پایگاه داده سنتی می‌باشد. با توجه به قابلیت‌های این پایگاه داده می‌توان این‌گونه نتیجه گرفت که پایگاه‌های داده توزیع شده نسبتی با حجم داده ورودی ندارد. بنابراین رابطه‌ای که می‌توان برای پایگاه داده‌ی کلان داده در نظر گرفت به صورت P_D در فرمول ریاضی ۲ خواهد بود:

$$P_D = K \quad (۲)$$

در این رابطه متغیر K مقداری ثابت می‌باشد که بستگی به نوع پایگاه داده دارد. با توجه به نتایج فوق، کارایی ترکیب پایگاه داده سنتی و توزیع شده هدوپ برای کلان داده شامل پارامترهای مختلف نظیر ارزانی، مقیاس‌پذیری، تحمل، انعطاف، بخش‌پذیری، تنوع داده‌ای، حجم داده، دسترسی‌پذیری، ترمیم و همزمانی می‌باشد که در مجموع ۱۰ قابلیت کارایی را شامل می‌شود.

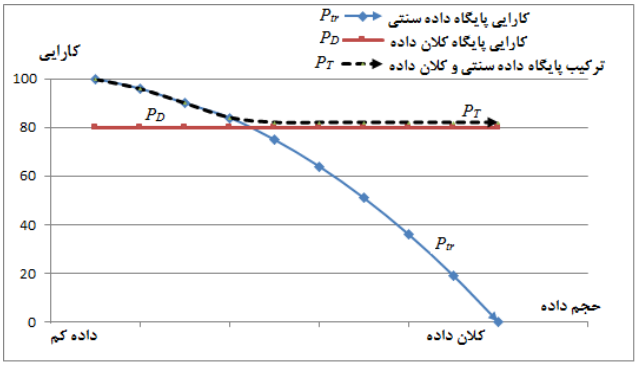
جدول ۳. مقایسه انواع پایگاه‌های داده برای داده‌های کلان

پایگاه داده / متریک	محیط استاندارد	محیط هدوپ	مبتنی‌بر کلید/مقدار	مبتنی‌بر ستون	مبتنی‌بر مستندات	مبتنی‌بر گراف	سنتی رابطه‌ای
مقیاس‌پذیری	✓		۴	۴	۳	۳	۱
		✓	۵	۵	۴	۴	
انعطاف‌پذیری	✓		۴	۳	۴	۴	۱
		✓	۵	۴	۴	۴	
پیچیدگی	✓		-۲	-۳	-۳	-۴	-۱
		✓	-۱	-۱	-۲	-۳	
سرعت اجرا	✓		۴	۴	۳	۲	۱
		✓	۵	۵	۴	۳	
حجم داده‌ها	✓		۴	۳	۳	۲	۱
		✓	۵	۴	۴	۳	
تنوع داده‌ای	✓		۴	۲	۳	۱	۱
		✓	۵	۳	۳	۳	
تحمل خطا	✓		۴	۴	۳	۳	۲
		✓	۵	۵	۴	۴	
دسترسی‌پذیری	✓		۴	۲	۴	۲	۴
		✓	۵	۳	۴	۳	
ثبات	✓		۲	۴	۲	۴	۴
		✓	۳	۵	۴	۴	
بخش‌بندی	✓		۴	۴	۴	۴	۱
		✓	۵	۵	۴	۴	
قابلیت اجرا در ابر	✓		۴	۴	۳	۳	۲
		✓	۵	۵	۵	۵	
مجموع امتیازات	✓		۳۶	۳۱	۲۹	۲۴	۱۶
		✓	۴۷	۴۳	۳۸	۳۴	
درصد امتیازات (قابلیت‌ها پایگاه داده)	✓		٪۷۲	٪۶۲	٪۵۸	٪۴۸	٪۳۲
		✓	٪۹۴	٪۸۶	٪۷۶	٪۶۸	

ترکیب پایگاه داده‌ی کلان داده‌ها و سنتی باعث افزایش کارایی سیستم به صورت ماگزیمم کارایی در ابعاد مختلف داده برای بازه‌های مختلف خواهد شد. بنابراین رابطه‌ای که می‌توان برای ترکیب پایگاه داده‌ها با توجه به جمیع مشخصات گفته شده، عنوان کرد به صورت ماگزیمم کارایی هر دو نوع پایگاه داده به صورت P_T در فرمول ریاضی ۳ است:

$$P_T = Max(P_{Tr} , P_D) \quad (۳)$$

این رابطه کارایی ترکیب پایگاه داده‌ی کلان داده‌ها و سنتی را نشان می‌دهد. جهت نشان دادن این ارتباط، حجم داده در محدوده داده کم و کلان داده در دو پایگاه داده SQL و هدوپ مورد ارزیابی قرار گرفته که نتایج کلی در شکل ۱۰ ارائه شده است. این نتایج کارایی ترکیب پایگاه داده سنتی و کلان داده را نشان می‌دهد. این شکل، مقایسه‌ای از کارایی سه حالت مختلف برای پایگاه داده سنتی، کلان و ترکیب هر دوی آنها را می‌باشد.



شکل ۱۰. مقایسه کارایی پایگاه داده سنتی، کلان داده و ترکیب آنها

همانطور که در این شکل نشان داده شده است افزایش حجم داده در وهله اول باعث کاهش کارایی در پایگاه داده سنتی گردیده بطوری که با افزایش میزان داده ورودی کارایی پایگاه داده بطور نمایی کاهش می‌یابد تا اینکه توانایی خودر را در ارائه داده‌های زیاد از دست می‌دهد. از طرف دیگر در پایگاه داده توزیع شده میزان حجم داده نقش مهمی در عمکرد سیستم ایجاد نمی‌نماید و افزایش حجم داده تغییر چندانی در سطح کارایی ایجاد نمی‌نماید. لذا همان‌طور که این شکل نشان می‌دهد در حجم داده‌های کم، پایگاه‌های داده سنتی و در حجم داده‌های زیاد، پایگاه‌های داده توزیع شده، برتری داشته که در حالت کلی رفتار حالت ترکیبی بیشترین کارایی را ارائه می‌نماید. با توجه به چالش اصلی تنوع و حجم که دنیای کلان داده با آن مواجه است به طوری که یک مجموعه عناصر داده‌ای هر یک باید در یکی از انواع پایگاه‌های داده ذخیره شوند و هیچ پایگاه داده واحدی برای ذخیره‌سازی همه آنها وجود ندارد. نکته مهمی که در راهکار پیشنهادی اهمیت دارد اینکه برای داده‌های با حجم بسیارکم، از پایگاههای داده سنتی SQL و برای حجم انبوه از پایگاه‌های داده توزیع شده هدوپ استفاده می‌شود. بدین ترتیب امکان پردازش بهینه آنها میسر می‌شود. این ویژگی لزوم ترکیب هر دو پایگاه داده را برای بهترین کارایی در شرایط مختلف ایجاد می‌نماید.

۱۰- نتیجه‌گیری

این مقاله ضمن معرفی پایگاه‌های داده موجود و چگونگی مشکلات و نقاط ضعف ابزارهای سنتی در رویارویی با نیازهای جدید و آتی فناوری اطلاعات، از جمله حجم بزرگ داده‌ها و ترکیب ساختاری آنها، به معرفی پایگاه‌های داده مناسب جهت ذخیره داده‌های کلان می‌پردازد. بدیهی است با گسترش روز افزون زیرساخت‌ها، سرویس‌ها و برنامه‌های کاربردی جدید درفناوری اطلاعات و تولید غیرقابل تصور داده‌ها، ذخیره‌سازی، جستجو، مدیریت، پردازش و ارائه داده‌ها و سرویس‌های مرتبط در ابعاد وسیع کشوری و جهانی به پدیده‌ای بسیار پیچیده و فنی تبدیل شده است. ارائه راه‌حل مناسب به این تنگناها نیاز به ابزارها و راه‌حل‌های هوشمند دارد. در این تحقیق جنبه‌های ساختاری و کارکردی پایگاه‌های داده سنتی مورد بررسی قرار گرفت و نشان داده شد که اصول ریاضی پشتیبان در این روش و مدل‌های پیاده‌سازی آن ویژگی‌های خاصی را ایجاد می‌نماید که مناسب داده‌های سنتی بوده و قابلیت حفظ و ذخیره‌سازی اطلاعات کم و متوسط را دارد. لذا با توسعه روزافزون صنعت فناوری اطلاعات و تغییر مدل و الگوهای تولید داده، پایگاه داده‌های سنتی در رویارویی با فضای جدید کارآیی خود را از دست داده‌اند. در این راستا با معرفی ویژگی‌های ساختاری و کارکردی پایگاه‌های داده غیررابطه‌ای، به نحوه ذخیره‌سازی کلان داده‌ها اشاره شده است. همچنین متد کلی جهت انتخاب پایگاه داده مناسب برای کلان داده‌ها با ترکیب پیوند پایگاه‌های داده سنتی و نوین جهت ذخیره و پردازش داده‌های حاصل از خدمات فراگیر ملی پیشنهاد شده است. علاوه بر آن روش‌هایی جهت مهاجرت از پایگاه‌های داده سنتی به پایگاه‌های داده کلان مورد بررسی قرار گرفتند. به منظور ارزیابی روش‌های ارائه شده، الگوهای کارایی شامل مقایسه کلی ویژگی‌های پایگاه‌های داده سنتی و کلان و همچنین مقایسه بر اساس متریک‌های انتخابی در بستر استاندارد و هدوپ با استفاده از روش کارت امتیازی متوازن مورد ارزیابی قرار گرفته است. نتایج بدست آمده نشان می‌دهد که ابعاد و ویژگی‌های داده تاثیر بسیار مهمی در انتخاب پایگاه داده مناسب داشته و در حالت کلی پایگاه داده با مدل ذخیره‌سازی کلید/مقدار بیشترین قابلیت برای ذخیره کلان داده‌ها را دارا می‌باشد. همچنین با در نظر گرفتن زیرساختار

هدوپ در بستر پایگاه داده، قابلیت‌های آن در تمامی متریک‌ها بهبود می‌یابند. از لحاظ کارایی، نتایج نشان می‌دهد که برای داده‌های با حجم کم، پایگاه‌های داده سنتی رابطه‌ای و در حجم داده‌های زیاد پایگاه‌های داده توزیع شده ارجحیت زیادی دارند. لذا در مجموعه داده‌های مختلف با حجم داده‌های متفاوت استفاده ترکیبی از هر دو نوع پایگاه داده بیشترین کارایی را به همراه خواهد داشت.

جهت تحقیقات بعدی، پیاده‌سازی روش پیشنهادی و تعیین جزئیات فنی در دسترسی به پایگاه‌های داده مختلف و تدوین راهکار اجرایی به منظور ذخیره‌سازی هر داده اختیاری می‌باشد.

مراجع

[1] Farahani, Farzane, and Fatemeh Rezaei. "Implementing a scalable data management system for collected data by smart meters." *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*. IEEE, 2021.

[2] ElDahshan, Kamal A., AbdAllah A. AlHabshy, and Gaber E Abutaleb. "A Comparative Study Among the Main Categories of NoSQL Databases." *Al-Azhar Bulletin of Science* 31.2 (2020): 51-60.

[3] de Oliveira, Vitor Furlan, et al. "SQL and NoSQL Databases in the Context of Industry 4.0." *Machines* 10.1 (2021): 20.

[4] Deka, G.C., *Handbook of Research on Securing Cloud-Based Databases with Biometric Applications*. 2014: IGI Global.

[5] Kausar, M.A. and M. Nasar, *SQL versus NoSQL databases to assess their appropriateness for big data application.* Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 2021. **14**(4): p. 1098-1108.

[6] Mostajabi, F., A.A. Safaei, and A. Sahafi, *A Systematic Review of Data Models for the Big Data Problems.* IEEE Access, 2021.

[7] Ramez, E. and N. Shamkant B, *Fundamentals of database systems*. 2022.

[8] Arora, I. and A. Gupta, *Cloud databases: a paradigm shift in databases.* International Journal of Computer Science Issues (IJCSI), 2012. **9**(4): p. 77.

[9] Ibrahem, Ali Hikmat, and Subhi RM Zeebaree. "Tackling the Challenges of Distributed Data Management in Cloud Computing-A Review of Approaches and Solutions." *International Journal of Intelligent Systems and Applications in Engineering* 12.15s (2024): 340-355.

[10] Strauch, S., et al. *Decision Support for the Migration of the Application Database Layer to the Cloud*. in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*. 2013. IEEE.

[11] Strauch, S., V. Andrikopoulos, and T. Bachmann. *Migrating application data to the cloud using cloud data*. in *e 3rd International Conference on Cloud Computing and Service Science,(CLOSER)*. 2013. Citeseer.

[۱۲] محمدرضا احمدی، داود ملکی و احسان آریانیان، کتاب پایگاه‌های داده پیشرفته (راهکار گذر به کلان داده‌ها و زنجیره‌های بلوکی). ناشر: اندیشه عصر، ۱۳۹۹.

[13] Date, C., *What Is Database Design, Anyway?* In: Database and Relational Theory (2019): 393-406.

[14] Adhikari, M. and S. Kar, NoSQL Databases. Handbook of Research on Securing Cloud-Based Databases with Biometric Applications, 2014: p. 109.

[15] Chaudhry, Natalia, and Muhammad Murtaza Yousaf. "Architectural assessment of NoSQL and NewSQL systems." *Distributed and Parallel Databases* 38.4 (2020): 881-926.

[16] Maricic, M., et al., Measuring the ict development: the fusion of biased and objective approach. Scientific Bulletin" Mircea cel Batran" Naval Academy, 2015. 18(2): p. 326.

[17] Odun-Ayo, Isaac, and Adeyemi Aina. "Development of a Cloud-Based Payroll Management System."

[18] Rafique, Ansar, et al. "CryptDICE: Distributed data protection system for secure cloud data storage and computation." *Information Systems* 96 (2021): 101671

[19] Feng, X., M. Conrad, and K. Hussein, NHS Big Data Intelligence on Blockchain Applications, in Big Data Intelligence for Smart Applications. 2022, Springer. p. 191-208.

[20] Pikuleva, N., A.S. Khafizova, and D. Gashigullin. Querying Big Graphs in Data Flow Language. in 2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). 2021. IEEE.

[21] Chang, W.L., A. Roy, and M. Underwood, NIST big data interoperability framework: volume 4, security and privacy. 2015.

[22] Rassam, Murad, Aishah Alfarhan, and Reem Alhussain. "Cloud Database Security Issues and Challenges: A Review." *Journal of Innovative Information and Communication Technology* 1.1 (2021): 21-31.

[23] Khan, Wisal, et al. "SQL and NoSQL database software architecture performance analysis and assessments—A systematic literature review." *Big Data and Cognitive Computing* 7.2 (2023): 97.

[24] Vera-Olivera, H., et al., Data Modeling and NoSQL Databases-A Systematic Mapping Review. ACM Computing Surveys (CSUR), 2021. 54(6): p. 1-26.

[25] Oliveira, C., et al., Using the balanced scorecard for strategic communication and performance management, in Strategic corporate communication in the digital age. 2021, Emerald Publishing Limited.

[26] Goldstein, James C. "Strategy maps: the middle management perspective." *Journal of Business Strategy* 43.1 (2022): 3-9.

[27] Mohapatra, H. and A.K. Rath, Fault tolerance in WSN through uniform load distribution function. International journal of sensors wireless communications and control, 2021. 11(4): p. 385-394.

[28] Adam, K., et al. Bigdata: Issues, challenges, technologies and methods. in Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015). 2019. Springer.

[۲۹] مهدی مرسلی، ابوالفضل طرقی حقیقت، ساسان حسینعلی زاده، مروری بر کاربرد الگوریتم‌های فراابتکاری در توازن بار در رایانش ابری، فصلنامه فناوری اطلاعات و ارتباطات ایران، شماره ۵۳، سال ۱۴، پاییز-زمستان ۱۴۰۱

[۳۰] کیومرث سلیمی، مهدی ملامطلبی، بهبود توازن بار در رایانش ابری با استفاده از الگوریتم جهش قورباغه سریع *(R-SFLA)*، فصلنامه فناوری اطلاعات و ارتباطات ایران، شماره ۵۷، سال ۱۵، پاییز ۱۴۰۲