

فصل هفتم

آشنایی با مفاهیم پیشرفته

هدف کلی

آشنایی با مفاهیم پیشرفته در پایگاهداده‌ها و اهمیت دانش نهفته در پایگاهداده‌های بزرگ

هدف‌های یادگیری

۱. خوانندگان گرامی با مفاهیم پیشرفته و جدید در پایگاهداده‌ها آشنا و به پژوهش در این زمینه ترغیب می‌شوند.
۲. ایجاد دید وسیع و جهانی مانند پایگاهداده‌های ابری و کلانداده‌ها و فضای مجازی یکی دیگر از اهداف این فصل است. خوانندگان گرامی با خواندن این فصل پایگاهداده‌ها را در افقی وسیع‌تر از یک کامپیوتر، یک آزمایشگاه و یک دانشگاه خواهند دید.
۳. خوداتکایی، کاربرد، و کارآفرینی از مهم‌ترین اهداف آموزش است. خوانندگان گرامی تشویق می‌شوند که بیشتر و بیشتر به کاربردهای روزافزون پایگاهداده‌ها بیندیشند و ایده‌های خود را بیازمایند.
۴. دانشجویان گرامی ترغیب می‌شوند که با تعریف پروژه پایانی کارشناسی خود در زمینه پایگاهداده‌های جدید، یک قدم به پیش بروند و یک سیستم نرم‌افزاری طراحی و پیاده‌سازی کنند.

مقدمه

دانشجویان گرامی دوره کارشناسی لازم است با مفاهیمی که در دوره‌های بالاتر تدریس می‌شود، آشنا شوند تا بتوانند در صورت لزوم شخصاً پیگیری کرده و مطالب مورد نیاز خود را بازیابند. در این فصل تعدادی از این مفاهیم را معرفی می‌کنیم، از شاخص تا کلان‌داده و محاسبات ابری، از پایگاهداده‌های غیرساخت‌یافته تا داده‌کاوی، و از پایگاهداده‌های توزیعی (نامتمرکز) تا سیّار. مفاهیم پیشرفته بسیار زیاد و متنوع هستند و در این مختصر، مجال پرداختن به همه آن‌ها نیست. آشنایی با مفاهیم پیشرفته در پایگاهداده‌ها هدف این فصل است تا دانشجویان گرامی به شرکت در دوره‌های بالاتر و نیز پژوهش و بازیابی شخصی و خوداتکایی ترغیب شوند. آینده شغلی تعدادی از فارغ‌التحصیلان به پایگاهداده مرتبط خواهد بود و باید بتوانند در این زمینه اعتماد به نفس لازم را داشته و به نیروهای کارآفرین تبدیل شوند.

۱-۷ شاخص

پرونده‌های شاخص دار حداقل دو بخش دارند: ناحیه اصلی و شاخص. دستیابی پی‌درپی در پرونده‌های شاخص دار بسیار سریع‌تر از پرونده‌های متوالی است ولی هدف اصلی آن‌ها دستیابی تصادفی (مستقیم) است. در اینجا نوع پرکاربرد آن، یعنی پرونده‌های درختی شاخص دار را معرفی می‌کنیم.

درخت به طور معجزه‌آسایی، سرعت ذخیره و بازیابی را بالا می‌برد. به این‌منظور، در ساختن درخت شاخص، ترتیب رکوردها منظور و در بازیابی رکوردها از این ترتیب استفاده می‌شود.

مثال ساده: درخت جستجوی دوتایی را درنظر بگیرید. در ساختن این درخت، در هر گره، قانونی چون «کوچک به چپ، در غیر این‌صورت به راست» رعایت و رکوردها درجای خود ذخیره می‌شوند. در بازیابی، از ریشه شروع و با همان قانون به سمت چپ یا راست می‌رود. سرعت بازیابی از این نظر بالا است که در هر مقایسه، نیمی از داده‌های باقی‌مانده کنار گذاشته می‌شوند و به سرعت به برگ‌ها می‌رسد. البته درخت جستجو لزومی ندارد دوتایی باشد.

امتیازهای زیر را می‌توان برای پروندهای درختی شاخص‌دار قائل شد:

- سرعت بازیابی بسیار بالا
- جایزبودن کلیدهای تکراری
- درخت با بزرگ و کوچک شدن پرونده تغییر می‌کند و هیچگاه به بن‌بست نمی‌رسد.
- امکان چند شاخص روی یک پرونده

نگهداری شاخص‌ها در حافظه اصلی یا حافظه پنهان: از آنجاکه شاخص و ناحیه اصلی دو بخش کاملاً مجزا هستند و اندازه شاخص به نسبت اصلی بسیار کوچک است، معمولاً می‌توان شاخص‌ها را در حافظه اصلی یا حافظه پنهان نگهداری کرد و سرعت جستجو را به مراتب افزایش داد.

ارزش واقعی درخت در ذخیره و بازیابی اطلاعات در پروندهای B+tree به نمایش گذاشته می‌شود. B+tree نه تنها پرکاربردترین نوع پرونده شاخص‌دار است، بلکه درمجموع یکی از پرکاربردترین انواع پرونده‌است. با داشتن B+tree می‌توان پرس‌وجوهای مختلفی را، اعم از دستیابی متوالی یا تصادفی روی انواع کلیدها پاسخ گفت. امتیاز بزرگ پروندهای درختی شاخص‌دار این است که می‌توان روی یک پرونده، چند شاخص تعریف کرد. این امتیاز در B+tree به اوچ خود می‌رسد، زیرا می‌توان همزمان چندین شاخص (کلید) را برای هر دو نوع دستیابی به کار گرفت.

۲-۷ کلانداده و رایانش ابری

کلانداده (داده‌های بزرگ، داده‌های انبوه) معادل data mass یا big data است و برای اشاره به داده‌های پر حجم که توسط سازمان‌های بزرگ ذخیره و تحلیل می‌شوند مورد استفاده قرار می‌گیرد [Hashem et. Al. 2015]. این مجموعه‌های داده گاهی به قدری بزرگ هستند که با ابزارهای ذخیره و بازیابی و پایگاه‌های داده سنتی و معمولی قابل مدیریت نیستند. این موضوع هر روز جذابیت و مقبولیت بیشتری پیدا می‌کند، زیرا با استفاده از حجم‌های بیشتری از داده‌ها، می‌توان تحلیل‌های بهتر و پیشرفته‌تری را برای مقاصد مختلف، از جمله مقاصد تجاری، پزشکی و امنیتی انجام داد و نتایج مناسب‌تری دریافت کرد. بیشتر تحلیل‌های مورد نیاز در پردازش‌هایی مانند هواشناسی، تحقیقات ژنتیک، فیزیک، زیست‌شناسی و محیطی، جست‌وجو در اینترنت، تحلیل‌های اقتصادی و

مالی و تجاری مورد استفاده قرار می‌گیرد. حجم داده‌های ذخیره شده در کلان‌داده، عموماً به خاطر تولید و جمع‌آوری داده‌ها از مجموعه بزرگی از تجهیزات و ابزارهای مختلف مانند گوشی‌های موبایل، حسگرهای محیطی، لاغ نرم‌افزارهای مختلف، دوربین‌ها، میکروفون‌ها، شبکه‌های حسگر بی‌سیم و غیره با سرعت خیره‌کننده‌ای در حال افزایش است.

در کلان‌داده، مدیریت صحیح داده‌ها به منظور استخراج اطلاعات، کشف دانش و در نهایت تصمیم‌گیری درخصوص مسائل مختلف کاربردی شامل پنج فعالیت اصلی است:

۱. جمع‌آوری
۲. ذخیره‌سازی
۳. جستجو
۴. به اشتراک‌گذاری
۵. تحلیل

چالش‌های زیادی در حوزه کلان‌داده مطرح شده‌است که ابعاد مختلفی از مشکلات و ویژگی‌های این حوزه را بیان می‌کنند. این ویژگی‌ها در زیر خلاصه شده‌است:

- حجم داده: حجم داده‌ها به صورت نمایی در حال رشد است. منابع مختلفی نظیر شبکه‌های اجتماعی، لاغ سرورهای وب، جریان‌های ترافیک، تصاویر ماهواره‌ای، جریان‌های صوتی، تراکنش‌های بانکی، محتوای صفحات وب، اسناد دولتی و... وجود دارد که حجم داده بسیار زیادی تولید می‌کنند.
- سرعت یا نرخ تولید: داده‌ها از طریق برنامه‌های کاربردی و سنسورهای بسیار زیادی که در محیط وجود دارند با سرعت بسیار زیاد و به صورت بلاذرنگ تولید می‌شوند. بسیاری از کاربردها نیاز دارند به محضور ورود داده به درخواست کاربر پاسخ دهند. ممکن است در برخی موارد نتوانیم به اندازه کافی صبر کنیم تا مثلاً یک گزارش در سیستم برای مدت طولانی پردازش شود.
- تنوع: انواع منابع داده و تنوع در نوع داده بسیار زیاد است. درنتیجه ساختارهای داده‌ای بسیار زیادی وجود دارد. مثلاً در وب، افراد از نرم‌افزارها و مرورگرهای مختلفی برای ارسال اطلاعات استفاده می‌کنند. بسیاری از اطلاعات مستقیماً از انسان دریافت می‌شود و بنابراین وجود خطای اجتناب‌ناپذیر است. این تنوع سبب می‌شود

که جامعیت داده تحت تأثیر قرار بگیرد، زیرا هرچه تنوع بیشتری وجود داشته باشد، احتمال بروز خطای بیشتری نیز وجود خواهد داشت.

- صحت: با توجه به اینکه داده‌ها از منابع مختلف دریافت می‌شوند، ممکن است نتوان به همه آن‌ها اعتماد کرد. مثلاً در یک شبکه اجتماعی، ممکن است نظرهای زیادی درخصوص یک موضوع خاص ارائه شود؛ اما اینکه آیا همه آن‌ها صحیح و قابل اطمینان هستند، موضوعی است که نمی‌توان به سادگی از کنار آن گذشت. البته بعضی از تحقیقات این چالش را به معنای حفظ همه مشخصه‌های داده اصلی بیان کرده‌اند که باید حفظ شود تا بتوان کیفیت و صحت داده را تضمین کرد. در مولدهای کلان‌داده باید بتوان داده‌ای تولید کرد که نشان‌دهنده ویژگی‌های داده اصلی باشد.
- اعتبار: با فرض اینکه داده صحیح باشد، ممکن است برای برخی کاربردها مناسب نباشد یا به عبارت دیگر از اعتبار کافی برای استفاده در برخی از کاربردها برخوردار نباشد.
- نوسان: سرعت تغییر ارزش داده‌های مختلف در طول زمان می‌تواند متفاوت باشد. در یک سیستم معمولی تجارت الکترونیک، سرعت نوسان داده‌ها زیاد نیست و ممکن است داده‌های موجود مثلاً برای یک سال ارزش خود را حفظ کنند، اما در کاربردهایی نظیر تحلیل ارز و بورس، داده‌ها با نوسان زیادی مواجه هستند و به سرعت ارزش خود را از دست می‌دهند و مقادیر جدیدی به خود می‌گیرند. اگرچه نگهداری اطلاعات در زمان طولانی به منظور تحلیل تغییرات و نوسان داده‌ها حائز اهمیت است، افزایش دوره نگهداری اطلاعات، مسلماً هزینه‌های پیاده‌سازی زیادی را در بر خواهد داشت که باید در نظر گرفته شود.
- نمایش: یکی از کارهای مشکل در حوزه کلان‌داده، نمایش است. اینکه بخواهیم کاری کنیم که حجم عظیم اطلاعات با ارتباطات پیچیده، به خوبی قابل فهم و قابل مطالعه باشد، از طریق روش‌های تحلیلی و بصری‌سازی مناسب اطلاعات امکان‌پذیر است.
- ارزش: این موضوع دلالت بر این دارد که برای تصمیم‌گیری چقدر داده حائز ارزش است؛ به عبارت دیگر آیا هزینه‌ای که برای نگهداری داده‌ها و پردازش آن‌ها می‌شود، ارزش آن را از نظر تصمیم‌گیری دارد یا نه. معمولاً داده‌ها می‌توانند در لایه‌های مختلف جایه‌جا شوند. لایه‌های بالاتر به معنای ارزش بیشتر داده می‌باشند؛ بنابراین برخی از سازمان‌ها می‌توانند هزینه بالای نگهداری مربوط به لایه‌های بالاتر را قبول کنند.

رایانش ابری، مدل رایانشی بر پایه شبکه‌های رایانه‌ای مانند اینترنت است که الگویی برای عرضه و تحويل خدمات رایانشی (شامل زیرساخت، نرم‌افزار، بستر و سایر منابع رایانشی) با به کارگیری شبکه ارائه می‌کند. رایانش ابری از ترکیب دو کلمه رایانش و ابر ایجاد شده است. ابر در اینجا اشاره به شبکه‌های وسیع مانند اینترنت است که کاربر معمولی از پشت‌صفحه و آنچه در پی آن اتفاق می‌افتد، اطلاع دقیقی ندارد. رایانش ابری راهکارهایی برای ارائه خدمات فناوری اطلاعات به شیوه‌های مشابه با صنایع همگانی (آب، برق، تلفن و غیره) پیشنهاد می‌کند. این بدین معنی است که دسترسی به منابع فناوری اطلاعات در زمان تقاضا و بر اساس میزان تقاضای کاربر به‌گونه‌ای انعطاف‌پذیر و مقیاس‌پذیر از راه اینترنت به کاربر تحويل داده می‌شود.

تعریف رسمی متفاوتی در خصوص رایانش ابری می‌توان یافت. موسسه ملی فناوری و استانداردها رایانش ابری را این‌گونه تعریف می‌کند: «رایانش ابری مدلی است برای فراهم‌کردن دسترسی آسان بر اساس تقاضای کاربر از طریق شبکه به مجموعه‌ای از منابع رایانشی قابل تغییر و پیکربندی (مثل: شبکه‌ها، سرورها، فضای ذخیره‌سازی، برنامه‌های کاربردی و سرویس‌ها) که این دسترسی بتواند با کمترین نیاز به مدیریت منابع و یا نیاز به دخالت مستقیم فراهم‌کننده سرویس به سرعت فراهم شود».

کلان‌داده‌ها و محاسبات ابری به عنوان تکنولوژی‌هایی معرفی شده‌اند که بیشترین سرعت رشد را داشته‌اند. محاسبات ابری، به عنوان نمونه جدیدی برای ارائه زیرساخت‌های محاسباتی و پردازش داده‌های بسیار بزرگ برای انواع منابع به کار گرفته شده است. کلان‌داده‌ها با حجم بسیار زیاد، سرعت بسیار زیاد و تنوع بسیار زیاد به فرم‌های پردازشی جدیدی نیاز دارند که توانایی این را داشته باشند تا قادر به تصمیم‌گیری، کشف و بهینه‌سازی پردازش را افزایش دهند. برای اطلاعات بیشتر به [Imran Alan et. al.] مراجعه شود.

۳-۷ پایگاهداده NoSQL

عبارت NoSQL برای اشاره به پایگاهداده‌های متن باز به کار گرفته می‌شود که از رابط SQL استفاده نمی‌کنند و با پایگاهداده رابطه‌ای سنتی متفاوت هستند [Stonebraker, M. 2010]. این پایگاه‌های داده اغلب با مفاهیم سنتی نظیر جدول‌ها و سطر و ستون‌های ثابت بیگانه هستند و در بیشتر موارد، عملیاتی مانند پیوند در آن‌ها بی‌معنی است. امروزه تعداد زیادی از

آنها به بلوغ نسبی رسیده و پتانسیل بسیاری برای استفاده و توسعه در محیط‌های کلان‌داده را دارند. این پایگاه‌های داده جدید با عنوان NoSQL یا Not only SQL شناخته می‌شوند. عبارت NoSQL در حال حاضر برای اشاره به مجموعه‌ای از پایگاه‌های داده‌ای غیررابطه‌ای و گسترده در شبکه استفاده می‌شود که در نقطه مقابل پایگاه‌های رابطه‌ای سنتی مانند Oracle RDBMS، Oracle RDB، Microsoft SQL Server، Informix، MySQL، PostgreSQL قرار دارد.

همچنین کار روی پروژه‌ای به نام UnQL آغاز شده‌است که هدف آن، تدوین استاندارد زبان پرس‌وجو در پایگاه‌داده‌های NoSQL است. این زبان به‌گونه‌ای طراحی می‌شود که مجموعه‌ها، اسناد و فیلد‌های مشخصی را نیز مورد پرسش قرار دهد. این زبان فراتر از SQL به‌شمار می‌آید و زبان SQL را می‌توان نمونه محدودشده آن به‌شمار آورد. این زبان نیازی به تعریف طرح خاصی برای پایگاه‌داده ندارد. این ویژگی انعطاف‌پذیری بسیار بالایی به کاربردها می‌دهد که نیاز اساسی هر کسب‌وکاری است. پایگاه‌داده NoSQL به صورت خودکار داده‌ها را بدون نیاز به دخالت برنامه بر روی سرورها منتقل می‌کند و قدرت پرسش توزیعی از هزاران سرور و نیز cache کردن داده‌ها را دارد. کلان‌داده یکی از بهترین کاربردهایی است که پایگاه‌های داده NoSQL شایستگی خود را در فراهم کردن آن به اثبات رسانده‌است.

در این قسمت نمونه‌هایی از پایگاه‌های کلان‌داده متداول معرفی می‌شود:

- Amazon SimpleDB: این پایگاه‌داده غیررابطه‌ای بسیار دسترس‌پذیر، مقیاس‌پذیر و انعطاف‌پذیر است و به کاربر اجازه می‌دهد قبل از ذخیره‌سازی داده‌ها آنها را رمزگذاری کنند.

- Google App Engine Bigtable: این پایگاه‌داده یک سیستم ذخیره‌سازی توزیع شده برای داده‌های ساخت‌یافته است. علاوه‌بر آن در بسیاری از محصولات گوگل مانند Google app engine به خوبی پیاده‌سازی شده‌است و اجازه می‌دهد تا داده‌های پیچیده ذخیره شوند.

- Hadoop MapReduce: یک زیرساخت برنامه‌نویسی برای پیاده‌سازی مدل سیستم‌های توزیع شده و بیشتر مناسب داده‌های بدون ساختار است.

- CouchDB: یک پایگاه‌داده سندگرا است و به عنوان یک پایگاه‌داده شی‌گرا طراحی شده‌است. این پایگاه‌داده سرعت و مقیاس‌پذیری بالایی دارد.

۴-۷ داده‌کاوی

پایگاه‌های داده دربر گیرنده میلیاردها رکورد اطلاعاتی هستند و کشف و استخراج سریع و دقیق دانش از آنها ضروری است. در تمام کشورهای دنیا پایگاه‌های داده از ثبت احوال، کارمندان، دانشجویان و هزاران نمونه دیگر وجود دارد و می‌توان با آنالیز این اطلاعات به نتایج بسیار مهمی دست پیدا کرد که مدیران و تصمیم‌گیران جامعه را به‌سوی تصمیمات درست رهنمون کند. با توجه به این نیاز، بحث داده‌کاوی^۱ مطرح می‌شود [k. Rexer et. al 2011]. تعاریف بسیاری از داده‌کاوی وجود دارد که از آن جمله می‌توان به موارد زیر اشاره کرد:

- داده‌کاوی یعنی استخراج اطلاعات نهان و یا الگوهای روابط مشخص در حجم زیادی از داده‌ها در یک یا چند پایگاه‌داده بزرگ.
- داده‌کاوی یعنی کاویدن از منابع عظیم داده تا اطلاعات گرانبهای پنهان شده در حجم انبوهی از اطلاعات سطحی را استخراج کند.
- داده‌کاوی فرایندی است که در آن دانش ذخیره‌شده به صورت ضمنی در پایگاه‌داده‌های بزرگ را بازنمایی می‌کند.
- داده‌کاوی عبارت است از استخراج اطلاعات قابل پیش‌بینی از بانک‌های اطلاعاتی بزرگ.

می‌توان نتیجه گرفت که در عصر حاضر که به عصر اطلاعات شهرت یافته است، داده‌کاوی یکی از ضروریات هر جامعه‌ای است و در شاخه‌های متفاوتی می‌تواند مورد استفاده قرار گیرد. مراحل داده‌کاوی را می‌توان به صورت زیر خلاصه کرد:

۱. داده‌پیرایی: در این مرحله داده‌های مغشوش و ناسازگار حذف می‌شوند.
۲. یکپارچه‌سازی داده‌ها: در این مرحله داده‌هایی که در چند منبع مختلف قرار دارند، تجمعی و یکپارچه می‌شوند.
۳. انتخاب داده‌ها: در این مرحله داده‌هایی که مرتبط به کاوش مورد نظر ما هستند، بازیابی و انتخاب می‌شوند.
۴. تبدیل داده‌ها: در این مرحله داده‌های بازیابی شده به قالبی که برای شروع داده‌کاوی مناسب است، تبدیل می‌شوند.

۵. کاوش: مرحله اساسی کار که در آن با روش‌های هوشمند، الگوهای داده‌ها از استخراج می‌شوند.

۶. ارزیابی الگوهای جالب نشان‌دهنده دانش

۷. ارائه دانش: تکنیک‌های مختلفی که برای نمایش دانش وجود دارد به کار گرفته و دانش برای کاربران ارائه می‌شود.

الگوریتم‌های داده‌کاوی را نیز می‌توان به صورت زیر خلاصه کرد:

- **الگوریتم وابستگی^۱**: نوعی آنالیز پیوندی است که برای تشخیص رفتار یک رویداد و یا یک پروسه خاص استفاده می‌شود. این الگوریتم از دو بخش شرط و نتیجه تشکیل شده است. بخش شرط یک آیتم خاص از اطلاعات را کشف و در بخش نتیجه آیتمی دیگر از اطلاعات را که وابسته به شرط است پیدا می‌کند. در واقع شرحی از دو کلمه «اگر» و «پس»، برای کشف رابطه‌های ناشناخته میان اطلاعات است. برای مثال، می‌توان گفت که «اگر» شخصی کاغذ بخرد «پس» به احتمال زیاد قلم هم خریداری می‌کند. الگوریتم وابستگی بسیار مفید است، زیرا توسط آن می‌توانیم رفتار خریداران و... را تجزیه و تحلیل و پیش‌بینی کنیم.

- **الگوریتم خوشه‌بندی^۲**: اطلاعاتی را که ویژگی‌های نزدیک بهم و مشابه دارند را در قطعه‌هایی جداگانه که به آن خوشه گفته می‌شود، قرار می‌دهد. به بیان دیگر خوشه‌بندی همان دسته‌بندی‌های ساده‌ای است که در کارهای روزانه انجام می‌دهیم. بخش‌بندی داده‌ها به گروه‌ها یا خوشه‌های معنادار به طوری که محتویات هر خوشه دارای ویژگی‌های مشابه باشند را خوشه‌بندی می‌گویند.

- **الگوریتم درخت تصمیم^۳**: این الگوریتم داده‌ها را به مجموعه‌های مشخص تقسیم می‌کند. هر مجموعه نیز به چندین زیرمجموعه از داده‌های کم و بیش همگن که ویژگی‌های قابل پیش‌بینی دارند، تقسیم می‌شود. برای مثال فرض کنید که اطلاعاتی از محصولات فروخته شده خود دارید. با بررسی این اطلاعات مشخص می‌شود که تعداد ۹ خرید از ۱۰ فروش محصول موبایل توسط افراد ۱۵ تا ۲۵ ساله انجام گرفته است و تنها یک فروش برای افراد بالای ۲۵ سال داشته‌اید. از این اطلاعات می‌توان نتیجه گرفت که سن مشتری نقش مهمی در فروش موبایل دارد.

1. Association Algorithm

2. Clustering Algorithm

3. Decision Trees Algorithm

- **الگوریتم رگرسیون خطی^۱:** تکنیکی آماری برای بررسی و مدل‌سازی روابط میان داده‌ها است. رگرسیون خطی از فرمول‌های مناسبی جهت محاسبه مقادیر A و B برای رسیدن به پیش‌بینی C استفاده می‌کند. رگرسیون در داده‌کاوی، تنوع دیگری از درخت تصمیم است به شکلی که به محاسبه یک ارتباط خطی میان متغیرهای وابسته و غیروابسته کمک می‌کند. محاسبه‌های انجام شده در پیش‌بینی‌ها کاربرد دارند.
- **الگوریتم بیز^۲:** این الگوریتم برپایه قضیه بیز برای مدل‌سازی پیش‌گویانه ارائه شده‌است. قضیه بیز از روشی برای دسته‌بندی پدیده‌ها برپایه احتمال وقوع یا عدم وقوع یک پدیده استفاده می‌کند و احتمال رخدادن یک پدیده محاسبه و دسته‌بندی می‌شود.
- **الگوریتم شبکه‌های عصبی^۳:** شبکه‌های عصبی از پرکاربردترین و عملی‌ترین روش‌های مدل‌سازی مسائل پیچیده و بزرگ شامل صدھا متغیر است. هر شبکه عصبی شامل یک لایه ورودی است که هر گره در این لایه معادل یکی از متغیرهای پیش‌بینی است. این الگوریتم برای تجزیه و تحلیل داده‌های پیچیده‌ای که انجام آن توسط سایر الگوریتم‌ها به سادگی انجام نمی‌گیرد، کاربرد دارد. الگوریتم شبکه‌های عصبی در مواردی چون بازاریابی و پیش‌بینی حرکت سهام، نوسان‌های نرخ ارز و یا سایر اطلاعات سیال مالی که دارای پیشینه هستند، پیشنهاد می‌شود.
- **الگوریتم رگرسیون منطقی یا لجستیک^۴:** یک روش آماری است برای مدل‌سازی‌هایی که نتایج دودویی دارند، در شرایطی چون «برد» یا «باخت» در یک مسابقه حذفی که دو حالت بیشتر ندارد. از این مدل برای به دست آوردن نتایج بهینه پیش‌بینی استفاده می‌شود. از رگرسیون لجستیک می‌توان به عنوان تنوع دیگری از الگوریتم شبکه‌های عصبی نام برد. رگرسیون منطقی یک مدل آماری رگرسیون برای متغیرهای دودویی است.
- **الگوریتم خوش‌بندی زنجیره‌ای^۵:** این الگوریتم شباهت زیادی به خوش‌بندی دارد اما برخلاف الگوریتم خوش‌بندی، خوش‌ها را برپایه یک مدل جستجو می‌کند و نه

-
1. Linear Regression Algorithm
 2. Bayes Algorithm
 3. Neural Network Algorithm
 4. Logistic Regression Algorithm
 5. Sequence Clustering Algorithm

بر اساس شباهت رکوردها، این مدل زنجیره‌ای رویدادها را بر اساس زنجیره مارکوف ایجاد می‌کند. در زنجیره مارکوف توزیع احتمال شرطی حالت بعدی تنها به حالت فعلی بستگی دارد و به وقایع قبل از آن وابسته نیست. زنجیره مارکوف ابتدا یک ماتریسی از ترکیب تمامی وضعیت‌های شدنی (امکان‌پذیر) ایجاد می‌کند و سپس در هر خانه ماتریس احتمالات حرکت از یک وضعیت به وضعیت دیگر را ثبت می‌کند. از طریق این احتمالات محاسباتی انجام می‌شود که درنتیجه آن یک مدل برپایه آن ایجاد می‌شود.

- **الگوریتم سری‌های زمانی¹:** روش سری‌های زمانی یکی دیگر از روش‌های پیش‌بینی است. این مدل نوعی الگوریتم رگرسیون است که برای پیش‌بینی مقادیر پیوسته مانند فروش محصولات در طول زمان استفاده می‌شود. در این الگوریتم می‌توان از خروجی یک سری، برپایه رفتار سری دیگر استفاده کرد.

۷-۵ سیستم مدیریت جریان‌داده‌ها

جریان‌داده‌ها با پایگاه‌داده‌ها تفاوت‌های اساسی دارد. داده‌ها می‌توانند به صورت جریان‌های پیوسته‌ای از المان‌های داده دریافت شوند و همان کارها (از قبیل پردازش پرسش و...) روی این جریان‌های داده انجام شوند. امروزه کاربردهای فراوانی وجود دارند که داده‌های آن‌ها به صورت رابطه‌های مانا نیست، بلکه به شکل جریان‌های گذرا از داده‌ها است. برای مثال می‌توان به کاربردهایی نظیر مهندسی ترافیک و نظارت شبکه، ثبت مکالمات تلفنی، امنیت شبکه، کاربردهای مالی، شبکه‌های حسگر و پروسه‌های تولیدی اشاره کرد.

یک جریان‌داده، دنباله‌ای پیوسته، نامحدود، سریع، متغیر با زمان و شاید غیرقابل پیش‌بینی از المان‌های داده است که همواره به انتهای آن افزوده می‌شوند. هر المان داده می‌تواند یک مقدار ساده (مثلاً یک عدد صحیح که حاصل مقدار خوانده شده توسط یک حسگر است) باشد یا یک تاپل از یک رابطه (مثلاً ثبت مکالمات تلفنی). برای حفظ ترتیب المان‌ها در این دنباله، معمولاً در کنار هر المان یک مهر زمانی قرار می‌گیرد.

متأسفانه DBMS‌های سنتی قابلیت پردازش جریان‌های داده با این ویژگی‌ها را ندارند و نیاز به سیستم خاصی برای مدیریت جریان‌داده داریم که DSMS نام دارد. جریان داده‌ها باید به صورت بر خط پردازش یا در صورت نیاز بایگانی شود و گرنه از بین خواهد رفت. از طرفی، با توجه به نامحدودبودن اندازه جریان داده و محدودیت فضای ذخیره‌سازی، نمی‌توان داده‌های زیادی را بایگانی کرد، بنابراین نیاز به الگوریتم‌های تک‌گذره برای پردازش جریان‌های داده وجود دارد.

عملگرهای جریان داده‌ها نباید انسدادی باشند (عملگری را انسدادی گوییم که نتواند اولین تاپل خروجی‌اش را تولید کند مگر اینکه آخرین تاپل ورودی‌اش را در اختیار داشته باشد). قابلیت تطبیق سیستم با تغییر شرایط (مثلًاً تغییر در نرخ ورود داده‌ها و...) و امکان اجرای هم‌رونده‌چند پرسش نیز از دیگر ویژگی‌هایی است که در یک DSMS باید مدنظر قرار گیرد. با توجه به ویژگی‌های مطرح شده، مهم‌ترین تفاوت‌های یک DSMS با DBMS با توان به صورت زیر بیان کرد:

۱. DBMS‌ها از رابطه‌های پایدار و مانا استفاده می‌کنند ولی DSMS‌ها از جریان‌های داده گذرا نیز استفاده می‌کنند.

۲. پرس‌وجوها در DBMS از نوع یک‌بار مصرف است ولی در DSMS‌ها ممکن است پیوسته نیز باشد.

۳. دسترسی به داده‌ها در DBMS به صورت تصادفی و در DSMS‌ها متوالی می‌باشد.

۴. در DBMS فرض بر کافی‌بودن فضای رسانه در مقایسه با محدودبودن حافظه اصلی مورد استفاده در DSMS‌ها است.

۵. در DBMS فقط وضعیت فعلی سیستم اهمیت دارد، ولی در DSMS ترتیب رسیدن و اطلاعات گذشته نیز مهم است.

۶. DSMS برخلاف DBMS، فعال است و می‌تواند نتایج را بدون اینکه از او خواسته شود، ارسال کند.

۷. نرخ تغییر داده‌ها در DBMS معمولاً پایین و در DSMS‌ها بسیار زیاد (در حد رسیدن چندین گیگابایت داده در هر ثانیه) است.

۸. کاربردهای DSMS، برخلاف DBMS، عموماً بلاذرنگ است.

۹. داده‌ها در DSMS ممکن است دقیق نباشند.

۱۰. در DBMS، طرح پرس‌وجو می‌تواند از قبل و به‌طور ثابت مشخص شده باشد، ولی در DSMS به‌دلیل غیرقابل پیش‌بینی بودن ویژگی‌های جریان‌داده‌ها، چنین نیست.
پرس‌وجوهایی که به یک DSMS ارائه می‌شوند را می‌توان به انواع زیر تقسیم‌بندی کرد:

۱. پرس‌وجوهای یک‌بار مصرف: پرس‌وجو در یک لحظه زمانی خاص ارزیابی می‌شود.
۲. پرس‌وجوهای پیوسته: به‌طور پیوسته (با رسیدن المان‌های داده) ارزیابی می‌شود. ممکن است نتایج درجایی ذخیره و با رسیدن داده جدید به‌روزرسانی شوند یا اینکه نتیجه، خودش به‌صورت یک جریان‌داده ارسال شود.

از جهت دیگر، پرس‌وجو می‌تواند از پیش‌تعریف‌شده یا ویژه باشد.

۱. پرس‌وجوی از پیش‌تعریف‌شده: قبل از شروع به دریافت جریان، پرس‌وجو ثبت می‌شود. این نوع پرس‌وجو اغلب از نوع پیوسته است.
۲. پرس‌وجوی ویژه: پس از شروع شدن جریان‌داده ثبت می‌شود. می‌تواند پیوسته یا یک‌بار مصرف باشد و ممکن است نیاز به داده‌هایی که قبلاً رسیده‌اند (از بین رفته‌اند)، داشته باشد.

انواع پاسخ‌ها در DSMS عبارت‌اند از:

- یک‌بار مصرف: پاسخ پرس‌وجو یک بار محاسبه شده و برگردانده می‌شود.
- مبتنی بر رویداد یا زمان: بر اساس وقوع رویدادی خاص و یا رسیدن به زمان به‌خصوصی جواب را برمی‌گرداند. مثلاً پیام هشدار نفوذ به شبکه در کاربردهای مدیریت (نظرارت^۱) شبکه.
- متناوب: به‌طور متناوب و در بازه‌های زمانی تعیین‌شده جواب را برمی‌گرداند.
- پیوسته: به‌طور پیوسته و مداوم جواب را برمی‌گرداند. این جواب می‌تواند در یک رابطه ذخیره شود یا اینکه خود به عنوان یک جریان مورد استفاده قرار گیرد.
- با توجه به ویژگی‌های جریان‌های داده، مهم‌ترین نکاتی را که در پردازش آن‌ها باید رعایت شوند، می‌توان به‌صورت زیر بیان کرد:
- بلادرنگ: زمان لازم برای پردازش هر داده از جریان‌داده‌ها باید بسیار کوتاه و ناچیز باشد و سیستم به‌صورت برخط و بلادرنگ کار کند.

- تک‌گذره: هر داده را حداکثر می‌توان یک‌بار بررسی کرد و پس از آن از بین می‌رود.
- محدودیت در اندازه حافظه: با توجه به نامتناهی و نامحدودبودن جریان داده، نمی‌توان کل آن را در حافظه محدودی که در اختیار داریم، ذخیره کرد. البته از این حافظه محدود برای ذخیره خلاصه داده‌های گذشته استفاده می‌شود و به نام خلاصه^۱ معروف است.
- جواب تقریبی^۲: جواب محاسبه شده ممکن است کاملاً دقیق نباشد، بلکه جوابی تقریبی به کاربر ارائه شود.
- برای پاسخ به هر پرس‌وجو روی جریان داده ورودی، طرح پرسش^۳ ایجاد می‌شود. هر عملگر با استفاده از خلاصه‌هایی که از داده‌های قبلی در اختیار دارد، پردازش مربوطه را روی داده‌های جریان‌های داده ورودی که در صفات‌های ورودی قرار گرفته‌اند، انجام داده و نتیجه را در صفت خروجی خود قرار می‌دهد. این اطلاعات می‌تواند توسط عملگر بعدی مورد استفاده قرار گیرد.

۷-۶ پایگاهداده توزیعی(نامتمرکز)

پایگاهداده توزیعی(نامتمرکز) مجموعه‌ای از چند بانک اطلاعات است که از نظر معنا مرتبط به هم هستند و روی یک شبکه کامپیوتری توزیع شده‌اند [حقجو، ۱۳۹۳]. سیستم مدیریت بانک اطلاعات نامتمرکز نرم‌افزاری است که امکان مدیریت پایگاه‌های نامتمرکز را فراهم می‌کند و نامتمرکزبودن سیستم را از دید کاربران پنهان می‌سازد. سیستمی که در آن کل اطلاعات در یک سایت قرار دارد و سایت‌های دیگر از طریق شبکه از آن بانک استفاده می‌کنند، یک بانک نامتمرکز نیست، زیرا اطلاعات توزیع نشده‌است. چنین سیستمی فقط یک بانک اطلاعات متمرکز روی شبکه است. بانک اطلاعات نامتمرکز را می‌توان از نظر گستردگی جغرافیایی به « محلی^۴» (LAD)، « منطقه‌ای^۵» (MAD) و « گسترده^۶» (WAD) طبقه‌بندی کرد. هر چه برعه گستردگی جغرافیایی بانک اطلاعات نامتمرکز افزوده می‌شود، تعداد آن کم می‌شود. بانک‌های

-
1. Synopsis
 2. Approximate Answer
 3. Query Plan
 4. Local Area Database
 5. Metropolitan Area Database
 6. Wide Area Database

محلی از یک کامپیوتر اصلی و تعدادی پایانه مختلف و کامپیوتر شخصی تشکیل شده و در درون یک ساختمان یا حداکثر یک مجموعه واقع هستند و مدیریت آنها نسبتاً ساده‌تر است. بانک‌های منطقه‌ای از طریق خطوط تلفن به هم متصل هستند. بانک‌های گسترده‌تر نیز از کامپیوترها و پایانه‌های متفاوتی تشکیل شده‌اند و به یک شبکه سراسری کشوری، یا جهانی متصل هستند.

در پایگاهداده‌های همگون، تمام سایت‌ها دارای ویژگی‌های سخت‌افزاری و نرم‌افزاری مشابه هستند و برای پردازش درخواست کاربر با دیگران مشارکت دارند. در پایگاهداده‌های ناهمگون، سایت‌های مختلف می‌توانند از نرم‌افزار، سخت‌افزار و یا شمای متمایزی استفاده و امکاناتی را برای مشارکت در انجام کارهای سایت‌های دیگر ارائه کنند.

در همگونی از جهت نرم‌افزار، دو جنبه اصلی عبارت‌اند از: سیستم عامل و سیستم مدیریت بانک اطلاعات. ساده‌ترین شکل ممکن زمانی است که همه بانک‌ها همگونی داشته باشند. مرحله بعدی حالتی است که همه چیز یکسان باشد به جز سیستم‌های مدیریت بانک اطلاعات که از یک مدل خاص ولی متفاوت باشند (مثلًاً همگی از نوع رابطه‌ای باشند ولی یکی Oracle، دیگری DB2 و... باشد). پیچیده‌ترین حالت زمانی است که همه چیز و از جمله سیستم‌های مدیریت پایگاهداده متفاوت باشند، مثلًاً یک سایت O2 (از نوع شئگرا) اجرا کند، دیگری Oracle و... در این حالت که در حال حاضر کمابیش در جهان وجود دارد، همه سیستم‌ها از واسط یکسانی پیروی می‌کنند و قادرند با هم ارتباط برقرار کنند.

توزيع داده‌ها شامل دو بخش است: تقسیم یعنی شکستن پایگاهداده و تقسیم آن به تکه‌هایی که در سایت‌های مجزا ذخیره می‌شوند و تخصیص تکه‌ها به سایت‌ها.

در پایگاه داده‌های نامتمرکز، معمولاً برای سرعت بالاتر در ذخیره و بازیابی اطلاعات، قابلیت اطمینان، تحمل پذیری خطا و... یک قلم داده را در بیش از یک سایت نگهداری می‌کنند. این کار تکرار داده نام دارد که علاوه‌بر افزونگی، مشکل به‌روز نگهداشت این تکرارها را نیز به‌دبیال دارد.

چنانچه داده در تمام سایت‌ها تکرار شده باشد، تکرار کامل نام دارد و اگر فقط در برخی از سایت‌ها تکرار شده باشد، تکرار جزئی نامیده می‌شود.

شفافیت به این معنی است که پیچیدگی‌ها و جنبه‌های مختلف مربوط به توزیع جغرافیایی بانک اطلاعات حتی‌الامکان از دید کاربران مخفی باشد و آن‌ها را آزار ندهد. وضع ایده‌آل این است که کاربران در ذخیره و بازیابی اطلاعات (پرسش) هیچ‌کاری با نامتمرکزبودن نداشته باشند. سیستم می‌تواند یک پرسش را تجزیه و بهینه کند، پاسخ‌ها را دریافت، ترکیب و تسلیم کاربر کند. به عبارت دیگر، شفافیت بدین معنی است که کاربر نباید هیچ محدودیتی که مربوط به توزیع پایگاهداده است را احساس کند و برای او فرقی نداشته باشد که آیا با یک بانک متمرکز کار می‌کند یا نامترکز. شفافیت جنبه‌های مختلفی دارد مانند شفافیت خرابی به معنی جایگزینی خودکار بانکی با بانک دیگری که دچار خرابی می‌شود، شفافیت کارایی به معنی انتخاب خودکار کم‌هزینه‌ترین و سریع‌ترین راه حل، شفافیت ناهمگونی به معنی حل خودکار تفاوت‌ها، شفافیت تکرار داده، شفافیت تقسیم، شفافیت محل داده‌ها و غیره.

منظور از قابلیت اطمینان، توانایی سیستم در پاسخ به درخواست‌ها در صورت وقوع خرابی است. با داشتن عناصر (سخت افزاری، نرم‌افزاری، داده و...) متعدد در سایت‌های گوناگون، انتظار می‌رود ضمن حل مشکل وابستگی به یک سایت، قابلیت اطمینان سیستم نیز افزایش یابد.

مهم‌ترین جنبه‌هایی که به بهبود کارایی در بانک‌های نامترکز کمک می‌کنند عبارت‌اند از:

- موازی‌سازی
- هم درون یک پرس‌وجو و هم بین چند پرس‌وجو می‌توان موازی‌سازی انجام داد. مثلاً یک پرس‌وجو را به چند زیرپرس‌وجو تقسیم کرده، آن‌ها را به‌طور موازی در سایت‌های مختلف اجرا می‌کنیم.
- محلی‌کردن داده‌ها به معنی نزدیک‌کردن داده‌ها به محل استفاده آن‌ها.
- پایگاه‌های داده که در یک محیط نامترکز همکاری می‌کنند، مایل نیستند آزادی عمل خود را از دست دهند. این خودمختاری جنبه‌های مختلفی دارد. مثلاً ممکن است مایل نباشند دیگران داده‌های آن‌ها را تغییر دهند یا همه وقت کامپیوتر آن‌ها را بگیرند و صرف کارهای سراسری کنند. بعضی از جنبه‌های خودمختاری غیرقابل گذشت هستند. مهم‌ترین اصولی که بانک‌ها معمولاً برای خود محفوظ نگه می‌دارند

عبارت‌اند از: مالکیت و مدیریت داخلی داده‌ها، عدم وابستگی به بانک‌های دیگر در انجام امور داخلی، کنترل داخلی عملیات و آزادی در اعمال روش‌های امنیت و جامعیت. نتیجه‌ای که از خود مختاری داخلی می‌توان گرفت این است که وجود یک سایت مرکزی و وابستگی بانک‌های دیگر به آن مطروح است، زیرا در صورت خرابی سایت مرکزی سایر بانک‌ها نیز مختل می‌مانند.

در معماری پایگاه‌های داده‌های نامت مرکز ساده، شمای مفهومی سراسری^۱ (GCS) از ابتدا وجود دارد و شماهای مفهومی محلی^۲ (LCS) در هر یک از بانک‌های عضو، از روی آن ساخته می‌شوند. به عنوان مثال اگر بخواهیم پایگاه‌های داده سراسری دانشگاه‌ها را با این روش طراحی کنیم، ابتدا باید شمای مفهومی سراسری که فعالیت تمام دانشگاه‌ها را دربر می‌گیرد، طراحی شود و سپس شماهای مفهومی داخلی از روی آن ساخته شوند. طبیعی است که مثلاً شمای مفهومی یک دانشگاه عام با یک دانشگاه صنعتی متفاوت خواهد شد.

طراحی یک سیستم پایگاه‌های داده نامت مرکز شامل طراحی شبکه مربوطه و نیز تصمیم‌گیری در خصوص توزیع داده‌ها و نرم‌افزارها در سایت‌های این شبکه است. دو استراتژی کلی در طراحی وجود دارد: طراحی بالا به پایین و طراحی پایین به بالا.

در طراحی بالا به پایین، پس از تحلیل نیازمندی‌ها، ابتدا شمای مفهومی سراسری را تعیین کرده، سپس دیدها (شماهای خارجی) و توزیع مناسب را طراحی می‌کنیم. عملکرد سیستم همواره تحت نظارت خواهد بود و در صورت نیاز، طراحی مجدد انجام می‌شود. طراحی بالا به پایین برای مواردی مناسب است که بخواهیم سیستم را از آغاز بنا کنیم. بسیاری از مواقع، پایگاه‌های داده از قبل وجود دارند و هدف از طراحی، ارتباط و تعامل آن‌ها در قالب یک پایگاه داده سراسری است. در این صورت از طراحی پایین به بالا استفاده می‌شود. در این روش از شماهای مفهومی محلی سایت‌ها شروع می‌کنیم و از اجتماع بخش‌هایی از آن‌ها، شمای مفهومی سراسری را به دست می‌آوریم. بنابراین از ابتدا شمای مفهومی سراسری وجود ندارد. طراحی پایین به بالا معمولاً برای سیستم‌های ناهمگون و خودمختار به کار می‌رود.

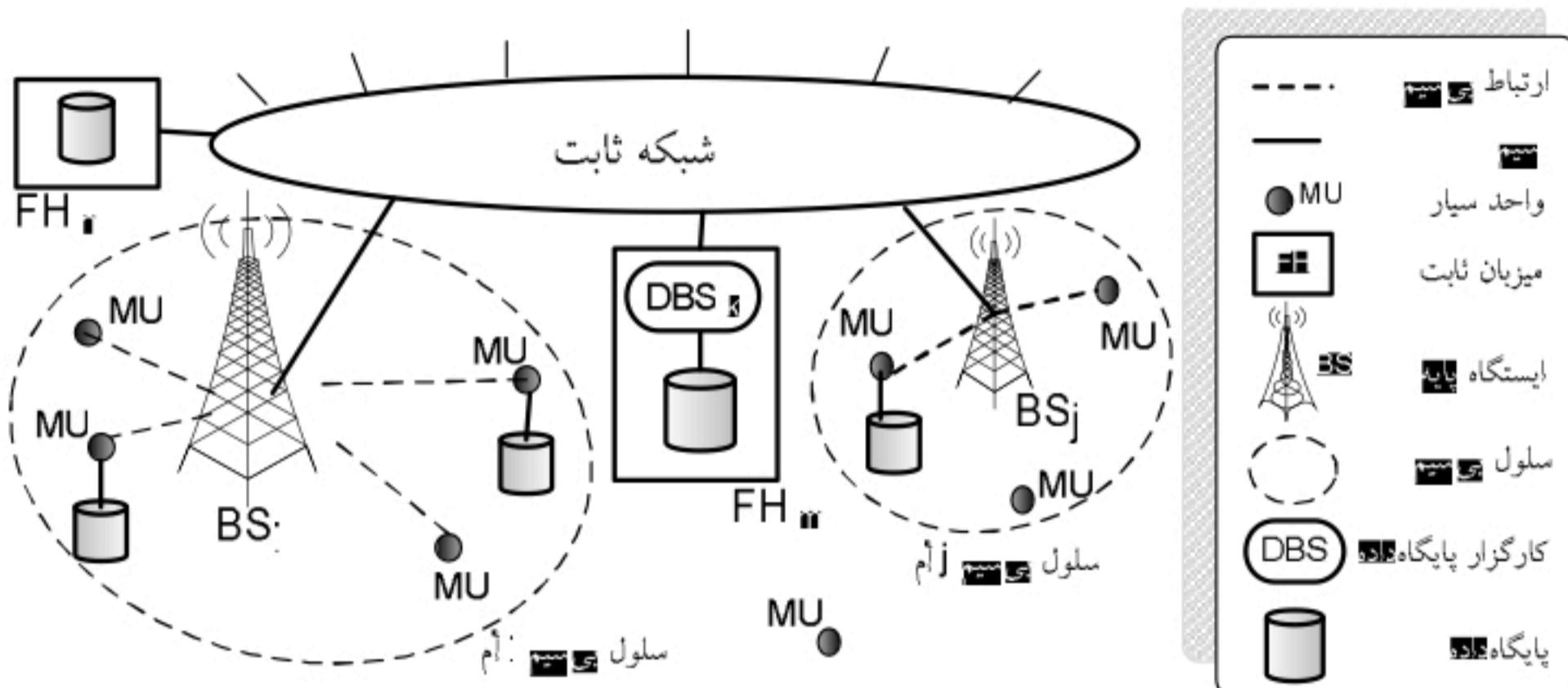
۷-۷ پایگاهداده سیار (موبایل)

امروزه دستیابی کاربران به اطلاعات از طریق موبایل،^۱ PDA، پخش‌های MP3 و یا کاربردهایی نظیر سیستم‌های ناوی بری در اتومبیل‌ها، به جزیی انکارناپذیر از زندگی روزمره تبدیل شده است که به آن رایانش سیار گفته می‌شود. سیار و قابل حمل بودن، گونه‌جدیدی از کاربردها را به ارمغان می‌آورد و پایگاهداده سیار در همین خصوص مطرح شده است که در آن تعداد زیادی از دستگاه‌های کامپیوتری از طریق کانال‌های ارتباطی بی‌سیم بر روی پایگاهداده، پرس‌وجو انجام می‌دهند و تراکنش اجرا می‌کنند [حق جو، ۱۳۹۳]:

سیستم پایگاهداده سیار، در حقیقت سیستم بانک اطلاعات نامتمرکزی است که محیط پردازشی آن سیار است و امکان محاسبات سیار را فراهم می‌کند. پایگاهداده ممکن است بر روی سایت‌های متعددی (چه واحدهای سیار و چه ثابت) ذخیره شده باشد. اجتماع پایگاهداده‌ها، کل سیستم پایگاهداده سیار را پدید می‌آورد. گره‌های پردازش گر ممکن است سیار باشند. هدف اصلی پایگاهداده سیار فراهم کردن اطلاعات مورد نیاز یک کاربر سیار است که ویژگی‌های زیر را دارد:

- توانایی تغییر موقعیت جغرافیایی: واحدهای سیار می‌توانند بدون اینکه اتصال پیوسته داشته باشند، در نواحی مختلف جغرافیایی تحت پوشش شبکه خود حرکت کنند.
 - قطعی مکرر: ممکن است اتصال یک واحد سیار از یک کارگزار قطع شود و همان لحظه به کارگزار دیگری متصل شود.
 - ارتباط بی‌سیم: یک واحد سیار می‌تواند با یک کارگزار و واحدهای سیار دیگر از طریق بی‌سیم ارتباط برقرار کرده و تبادل پیام کند.
 - توسعه‌پذیری: در هر زمان یک واحد می‌تواند به شبکه اضافه شود یا واحدهای موجود از شبکه حذف شوند.
 - ناهمگونی: سیستم می‌تواند شامل گونه‌های مختلفی از سخت‌افزار و انواع داده باشد و هر عضو ممکن است پایگاهداده و سیستم مدیریت پایگاهداده خود را داشته باشد.
- شکل ۱-۷ معماری عمومی پایگاهداده سیار را نشان می‌دهد.
- این معماری شامل اجزای زیر است:

- شبکه ثابت: یک شبکه متداول سیم کشی شده است که ارتباط اجزای ثابت را فراهم می کند.



شکل ۱-۷. معماری عمومی پایگاهداده سیار

- واحد سیار^۱ (MU): یک کامپیوتر سیار که قادر به برقراری ارتباطات بی سیم است.
- ایستگاه پایه^۲ (BS): یک فرستنده / گیرنده که مجهز به واسط بی سیم است. ارتباط بین شبکه ثابت و واحدهای سیار از طریق این واحد انجام می شود. فضای در محیط سیار به بخش هایی به نام سلول تقسیم می شود. هر سلول توسط یک ایستگاه پایه تحت پوشش قرار می گیرد. اندازه هر سلول با توجه به قدرت ایستگاه پایه آن تعیین می شود.
- کارگزار پایگاهداده^۳ (DBS): امکانات پایگاهداده را فراهم می آورد و وظیفه پردازش تراکنش ها و پاسخگویی به پرس و جوها را بر عهده دارد.
- میزبان ثابت: کامپیوتری که به شبکه ثابت متصل است و دارای ویژگی سیار بودن نیست. از آنجایی که به فرستنده / گیرنده مجهز نیست، قادر به ارتباط مستقیم با واحدهای سیار نیست.

برخی از مهم ترین ویژگی های محیط سیار با توجه به ویژگی های پایگاهداده سیار عبارت اند از:

- کمبودن پهنازی باند
- نامطمئن بودن کانال های بی سیم برای انتقال اطلاعات

1. Mobile Host (MH) or Mobile Unit (MU)
 2. Base Station
 3. Database Server (DBS)

- نامتقارن بودن^۱ ارتباطات: پنهانی باند درجهت روبه‌پایین جریان (از سرویس‌دهنده به مشتری‌ها) بسیار بیشتر از جهت معکوس آن است. اگرچه مشتری‌ها می‌توانند داده‌ها را با نرخ بالایی دریافت کنند، اما در برخی سیستم‌ها مشتری قادر به ارسال پیام‌های زیادی به سرویس‌دهنده نیست.
- قطع شدن متناوب: مشتری‌های سیار (برخلاف میزبان‌های ثابت) به‌طور پیوسته و مداوم به شبکه متصل نمی‌مانند، بلکه کاربران واحدهای سیار خود را به‌طور منظم خاموش و روشن می‌کنند. علاوه‌بر این مشتری‌های سیار می‌توانند از یک سلول جدا شده و به سلول دیگری ملحق شوند و به قولی در محیط پرسه بزنند.
- انرژی محدود: برخی از واحدهای قابل حمل از لحاظ میزان انرژی باتری (تا قبل از شارژ مجدد) به‌شدت محدودیت دارند.
- اندازه صفحه نمایش کوچک: بعضی از واحدهای سیار دارای صفحات نمایش بسیار کوچکی هستند که در کاربردها (مثلاً در نمایش جواب به کاربر) باید به این ویژگی‌ها نیز دقت کرد.
- پرسه‌زدن مشتری‌ها بین سلول‌ها نیز مسائلی مانند پرس‌وجوهای وابسته به مکان را به‌بار می‌آورد که می‌توان آن‌ها را چنین دسته‌بندی کرد:
 - پرس‌وجوهای «آگاه از مکان» مانند اسمی هتل‌های شهر تهران
 - پرس‌وجوهای «وابسته به مکان»: پاسخ پرس‌وجو وابسته به موقعیت مکانی صادرکننده^۲ آن است مانند شماره تلفن نزدیک‌ترین مکان

مثال: به عنوان نمونه‌هایی دیگر از پرس‌وجوهای وابسته به مکان به مثال‌های زیر توجه کنید:

 - هتل‌هایی که تا ۵ دقیقه دیگر به آن‌ها می‌رسم.
 - فرد X چقدر از اینجا دور است؟

خلاصه فصل هفتم

در این فصل با مفاهیم پیشرفتی پایگاهداده‌ها آشنا شدیم. مفاهیمی چون روش‌های بهینه ذخیره‌سازی پروندهای رایانش ابری و کلان‌داده‌ها، پایگاهداده NOSQL، داده‌کاوی و

1. Asymmetry
2. Query Issuer

الگوریتم‌های مختلف آن، سیستم مدیریت جریان داده‌ها، پایگاه‌داده‌های موازی و توزیعی و پایگاه‌داده‌های سیار را معرفی کردیم. هرچند بیشتر این مفاهیم در پایگاه‌داده‌ها سابقه چند ساله دارند، ولی برای دانشجویانی که در ابتدای این راه هستند، مفاهیم پیشرفته به حساب می‌آیند. دانشجویان عزیز به عمق و کارایی پایگاه‌داده‌ها در داده‌پردازی واقف می‌شوند و برای ادامه راه ترغیب می‌شوند.

تمرین‌های تشریحی فصل هفتم

۱. امکانات یک DBMS را با DSMS هم‌تراز آن مقایسه کنید (نقاط ضعف و قوت آن‌ها را در مقایسه با یکدیگر بنویسید).
۲. ذخیره‌سازی کلان‌داده در فضای ابر مناسب‌تر است یا کامپیوترهای معمول؟ چرا؟
۳. امتیاز ذخیره‌سازی کلان‌داده در فضای ابر چیست؟
۴. پایگاه‌داده نامت مرکزی مثال بزنید که همه گره‌های آن از یک نوع باشند. محل قرارگرفتن DBMS را با توجه به گره‌ها با رسم شکل نشان دهید.
۵. آیا بدون داشتن یک سرور ثابت می‌توان یک پایگاه‌داده سیار داشت؟ اگر همه ارتباط‌ها قطع شود، چگونه می‌توان وصل مجدد شد؟
۶. چرا تراکنش‌های ACID در محیط NO SQL پاسخ‌گو نیستند؟
۷. پایگاه دانش چه امتیازی بر پایگاه‌داده دارد؟ اگر بخواهیم یک پایگاه‌داده را به پایگاه دانش تبدیل کنیم، چه بخش‌هایی باید اضافه شود؟
۸. در پایگاه‌داده نشر کتاب چند پرسش مثال بزنید که از نوع دانش باشد و نتوان پاسخ آن را مستقیماً از پایگاه‌داده بیرون کشید (دانش نهفته).