به نام خدا

وزارت ارتباطات و فناوری اطلاعات
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
ITRC

پژوهشکده فناوری اطلاعات

"پروژه تدوین نقشه راه کلان داده‌ها"

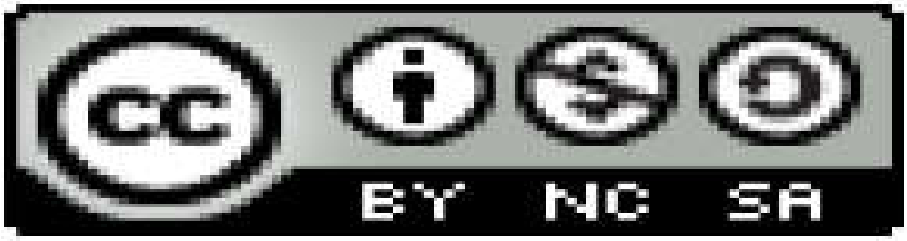
گزارش فاز اول

"سرویس‌های کلان‌داده‌ها"

کد پروژه: ۹۰۴۹۵۰۱۰۰

مجری:	محمدشهرام معین
تهیه‌کننده:	مسعود محمدزاده
کد گزارش:	
تاریخ ارائه:	۹۵/۱۲/۲۵
نسخه/وضعیت:	۲/نهایی

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد



گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

شناسنامه گزارش

عنوان: سرویس‌های کلان‌داده‌ها		**شماره نسخه:** Error! Reference source not found.
کد:	**نوع گزارش:** راهبردی	**تاریخ ارائه گزارش:** ۹۵/۱۲/۲۵
نام پروژه: تدوین نقشه راه کلان‌داده‌ها		**نوع پروژه: راهبردی**
تاریخ شروع: ۹۵/۷/۶	**تاریخ پایان:** ۹۶/۴/۶	
نام گروه: فناوری اطلاعات / سامانه های چندرسانه ای		
کد پروژه: ۹۰۴۹۵۰۱۰۰	**شماره و تاریخ قرارداد:** حکم شماره ۶۷۰۲/د/۵۰۰/پ در تاریخ ۹۵/۸/۱۰	
مجری: محمدشهرام معین	**ناظر / ناظرین:** آقایان دکتر علیرضا یاری، دکتر روح ا... رحمانی، دکتر مجید رسولی دیسفانی، دکتر امین شکری پور و مهندس فرزاد ابراهیمی	
تهیه کننده / تهیه کنندگان: مسعود محمدزاده		
نام و نشانی مجری: تهران، انتهای خیابان کارگر شمالی، پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران)_ کد پستی: ۱۴۳۹۹۵۵۴۷۱ _ تلفن: ۸۴۹۷۷۵۸۵		
نام و نشانی حمایت کننده: تهران، انتهای خیابان کارگر شمالی، پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران)_ کد پستی: ۱۴۳۹۹۵۵۴۷۱ _ تلفن: ۱۰-۸۰۰۵۵۰۸		
ملاحظات:		
چکیده: یکی از بخش‌های مهم نقشه راه کلان‌داده‌ها، برنامه‌ریزی برای ایجاد و گسترش سرویس‌های کلان‌داده است. این امر مستلزم شنا سایی سرویس‌هایی ا ست که در این حوزه باید ارائه شوند. برای شنا سایی این سرویس‌ها، گونه شنا سی سرویس‌های کلان‌داده‌ها ارائه شده و لایه‌های مختلف آن توضیح داده می‌شوند. سپس سرویس‌های موفقی که در سطح جهان در هرکدام از لایه‌های توضیح‌داده‌شده در گونه‌شناسی وجود دارند، معرفی شده و نحوه ارائه سرویس آنها تشریح می‌شود.		
کلمات کلیدی: کلان‌داده‌ها، سرویس ها، گونه شناسی		
وضعیت گزارش: نهایی		**زبان گزارش:** فارسی
وضعیت دسترسی: عادی		**تعداد صفحات:** ۳۰

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

چکیده

یکی از بخش‌های مهم نقشه راه کلان‌داده‌ها، برنامه‌ریزی برای ایجاد و گسترش سرویس‌های کلان‌داده است. این امر مستلزم شناسایی سرویس‌هایی است که در این حوزه باید ارائه شوند. برای شناسایی این سرویس‌ها، گونه‌شناسی سرویس‌های کلان‌داده‌ها ارائه شده و لایه‌های مختلف آن توضیح داده می‌شوند. سپس سرویس‌های موفقی که در سطح جهان در هرکدام از لایه‌های توضیح‌داده‌شده در گونه‌شناسی وجود دارند، معرفی شده و نحوه ارائه سرویس آنها تشریح می‌شود.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

اطلاعات مرتبط

مستندات مرتبط

شماره مستند	نوع مستند	نام مستند

تغییرات اعمال‌شده در نسخه‌های پیشین

شماره نسخه	تاریخ	تغییرات اعمال شده

تأییدکنندگان

	نام و نام خانوادگی	تاریخ	امضاء	ملاحظات
مجری پروژه	محمدشهرام معین			
تهیه کننده/ تهیه کنندگان	مسعود محمدزاده			
ناظر پروژه	دکتر علیرضا یاری، دکتر روح ا... رحمانی، دکتر امین شکری پور، دکتر مجید رسولی دیسفانی، مهندس فرزاد ابراهیمی			
مدیر گروه	مهندس فرزاد ابراهیمی			
مسئول مستندات پژوهشکده	مانا روزی طلب			
رئیس پژوهشکده/ معاون پژوهشی	دکتر علیرضا یاری / دکتر کامبیز بدیع			

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

تقدیر و تشکر

بدین وسیله از ناظرین و مشاورین محترم پروژه قدردانی می‌شود.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

سرفصل مطالب

۱ مقدمه ۱۰

۲ مفاهیم ۱۲

۳ گونه‌شناسی ۱۳

۳-۱ سرویس‌های داخلی ۱۴

۳-۱-۱ سرویس‌های هماهنگ‌کننده سامانه ۱۴

۳-۱-۲ سرویس‌های فراهم‌کننده داده‌ها ۱۵

۳-۱-۳ سرویس‌های فراهم‌کننده کاربرد کلان‌داده‌ها ۱۶

۳-۱-۴ سرویس‌های ارائه‌شده به مصرف‌کننده داده‌ها ۱۶

۳-۲ سرویس‌های عمومی ۱۷

۳-۲-۱ زیرساخت ابری به‌عنوان سرویس ۱۷

۳-۲-۲ مدیریت داده‌ها به‌عنوان سرویس ۱۹

۳-۲-۳ پلتفرم داده‌ها به‌عنوان سرویس ۲۰

۳-۲-۴ نرم‌افزار تحلیل به‌عنوان سرویس ۲۰

۴ سرویس‌های موفق کلان‌داده‌ها در جهان ۲۲

۴-۱ آمازون ۲۲

۴-۱-۱ EC2 ۲۲

۴-۱-۲ Dynamo ۲۴

۴-۱-۳ Elastic MapReduce ۲۴

۴-۲ گوگل BigQuery ۲۵

۴-۳ Splunk Storm ۲۶

۴-۴ مایکروسافت Azure-HDInsight ۲۶

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۴-۵ Tibco Silver Spotfire ۲۷
۴-۶ QuBole ۲۸
۵ جمع‌بندی ۲۹
مراجع ۳۱

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

فهرست اشکال

شکل ۱- سرویس‌های کلان‌داده‌ها ۱۳

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۱ مقدمه

یکی از وظایف مهم در حوزه ارتباطات و فناوری اطلاعات، تعریف واژگانی است که برای توصیف موارد مختلف استفاده می‌شوند. برای مثال، «محیط ابری» واژه‌ای است که در این حوزه زیاد استفاده می‌شود ولی بسیاری از افراد معانی مختلفی از آن ارائه می‌دهند. با افزایش واژگانی که معانی آنها به‌هم نزدیک بوده و به‌نوعی هم‌پوشانی دارند، این تعریف اهمیت دوچندانی نیز پیدا می‌کند. یکی از این موارد، واژگان «کاربرد»، «پلتفرم» و «سرویس» هستند که عدم وجود معانی دقیق برای آنها در بسیاری از موارد باعث ایجاد ابهام و سوءتفاهم می‌شود. بنا به تعریف، «کاربرد» برنامه‌ایست که برای انجام یک یا مجموعه‌ای عملکرد مربوط به‌هم موردنظر کاربر نهایی، طراحی شده است. همچنین، «سرویس» مکانیزمی برای تحویل‌دهی یک کاربرد است. اما معنی واژه «سرویس» در طول سال‌های گذشته از انتقال ساده داده یا تحویل‌دهی اطلاعات از یک نقطه به نقطه دیگر به مدلی برای دسترسی به کاربردی که به‌صورت یک سرویس، تحویل داده می‌شود، تکامل پیدا کرده است. سرویس‌ها معمولاً انعطاف‌پذیری کمتری نسبت به کاربردها دارند و عملکرد محدودتری را در اختیار کاربر قرار می‌دهند. همچنین در اغلب موارد، سرویس‌ها توسط کاربردهای دیگر مورداستفاده قرار می‌گیرند در حالی که همانگونه که در بالا نیز ذکر شد، کاربران «کاربردها» معمولاً کاربران نهایی هستند. معنی «پلتفرم» نیز عبارتست از سخت‌افزار، سیستم‌عامل یا نرم‌افزاری که «کاربرد» روی آن اجرا می‌شود. اما در رایانش ابری «پلتفرم به‌عنوان سرویس»، پلتفرم به معنی عملکردهای آماده قابل استفاده در قالب سرویس توسط توسعه‌دهندگان کاربرد است که توسعه‌دهنده آنها را میزبانی نمی‌کند ولی از طریق ارتباط اینترنتی آنها را به‌هم متصل می‌کند[۱] [۲]. بنابراین با توجه به تعاریف فوق می‌توان این نتیجه‌گیری را کرد که هر «کاربرد» یا «پلتفرمی» که از طریق اینترنت یا هر کانال دیگر انتقال داده در اختیار کاربر - که می‌تواند یک شرکت دیگر یا کاربر نهایی باشد - قرار گیرد، سرویس محسوب می‌شود. اما از آنجا که کاربردهای کلان‌داده‌ها در یک گزارش دیگر به‌صورت مفصل توضیح داده شده‌اند، در این گزارش به پلتفرم‌های ارائه‌شده به‌صورت سرویس که در معماری کلان‌داده مورداستفاده قرار می‌گیرند، اکتفا می‌کنیم.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

متخصصین سرشناس صنعت پیش‌بینی کرده‌اند که با به‌وجود آمدن پدیده کلان‌داده‌ها، در سال‌های آینده تقاضا برای نیروی متخصص در حوزه تحلیل کلان‌داده‌ها افزایش خواهد یافت. طبق این پیش‌بینی‌ها، در سال ۲۰۱۸ شکاف بین تقاضا و نیروی‌های متخصص موجود به ۱/۷ میلیون نفر خواهد رسید[۳]. تربیت نیروی انسانی در این حوزه محدود است و در برخی از اوقات، سازمان‌های متقاضی تمایل به تربیت نیروی متخصص توسط خود را ندارند. از آنجا که فناوری‌های سرویس‌های کلان‌داده‌ها (BDaaS) توسط فراهم‌کننده سرویس مدیریت می‌شود، این سرویس‌ها می‌توانند به سازمان‌ها در رسیدن به نیازهای خود در زمینه مهارت‌های لازم در فناوری‌های کلان‌داده‌ها کمک کنند. برای مثال اگر سازمانی به داشتن یک سامانه توزیع‌شده ذخیره‌سازی داده‌ها مانند HDFS نیاز داشته باشد، باید متخصصین زیرساخت را که آن دانش و مهارت‌های خاص را دارد، استخدام کند. نگه‌داری زیرساخت کلان‌داده‌ها جدید نیز هزینه مضاعفی را به خاطر نیاز به تخصص‌های گران‌قیمت که باید جذب شوند، به آن سازمان تحمیل خواهد کرد. پیاده‌سازی کلان‌داده‌ها برای سازمان‌های بزرگ نیز پرهزینه است زیر متخصصینی که دارای مهارت‌های کافی در این مورد هستند، بسیار محدودند. هزینه پیاده‌سازی پروژه‌های کلان‌داده‌ها، ۲۰ برابر نرم‌افزار مربوطه تخمین زده شده است[۴].

با استفاده از فناوری‌های ارائه‌شده در قالب سرویس‌های کلان‌داده‌ها که زیرساخت‌های لازم در این حوزه را فراهم می‌سازند، سازمان‌ها می‌توانند نیازهای کلان‌داده‌ها خود را بدون نیاز به سرمایه‌گذاری زیاد و استخدام نیروهای متخصص گران‌قیمت، برآورده سازند. در این روش، شرکت‌های فراهم‌کننده سرویس‌های کلان‌داده‌ها، نگه‌داری و تحقیق و توسعه فناوری‌های کلان‌داده‌ها را به‌عهده خواهند داشت و بنابراین سازمان‌های سرویس‌گیرنده از آنها می‌توانند روی برآورده کردن نیازمندی‌های کسب‌وکار خود با استفاده از کلان‌داده‌ها تمرکز کنند.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۲ مفاهیم

کلان‌داده‌ها: دادگانی هستند که قابل ذخیره‌سازی و پردازش با معماری‌های معمول نیستند. چهار مشخصه حجم، تنوع منبع، نرخ تولید و تغییرات، تعیین‌کننده معماری جدیدی هستند که برای اینگونه داده‌ها باید به‌کار گرفته شود.

سرویس کلان‌داده: سرویسی است که یک شرکت ثالث به شرکت اصلی می‌دهد و طی آن کل یا بخشی از روند جمع‌آوری، تحلیل، ذخیره‌سازی و بازارائه کلان‌داده‌ها را برعهده می‌گیرد.

سرویس ابری: سرویس‌هایی هستند که قابلیت‌هایی از قبیل رایانش و ذخیره‌سازی را در یک محیط ابری در اختیار کاربران قرار می‌دهند.

سامانه عمودی: سامانه‌هایی هستند که سرویس‌های محدود را برای کاربرد در یک حوزه خاص ارائه می‌دهند.

هماهنگ‌کننده سامانه: واحدی از سامانه عمودی کلان‌داده‌هاست که یکپارچه‌سازی فعالیت‌های سایر واحدهای سامانه را انجام می‌دهد.

فراهم‌کننده داده‌ها: یکی از بازیگران اکوسیستم کلان‌داده‌ها هستند که وظیفه تجمیع داده‌های خام یا داده‌های تحلیل‌شده توسط سایر سامانه‌های کلان‌داده‌ها را بر عهده دارند.

فراهم‌کننده کاربرد: یکی از بازیگران اکوسیستم کلان‌داده‌ها هستند که وظیفه آماده‌سازی، تحلیل و بصری‌سازی داده‌ها و ارائه آنها از طریق سرویس یا رابط کاربری به کاربر را دارد.

سامانه پایگاه داده رابطه‌ای (RDS-SQL): سامانه پایگاه داده‌ای است که مدل ذخیره‌سازی و مدیریت داده‌ها در آن به‌صورت جدول‌های مرتبط باهم است تا افزونگی داده‌های ذخیره‌شده به حداقل برسد.

سامانه پایگاه داده غیر رابطه‌ای (NoSQL): سامانه پایگاه داده‌ای است که قوانین رابطه‌ای حاکم بر سامانه‌های SQL را برای ذخیره‌سازی و دستکاری داده‌ها دنبال نمی‌کنند.

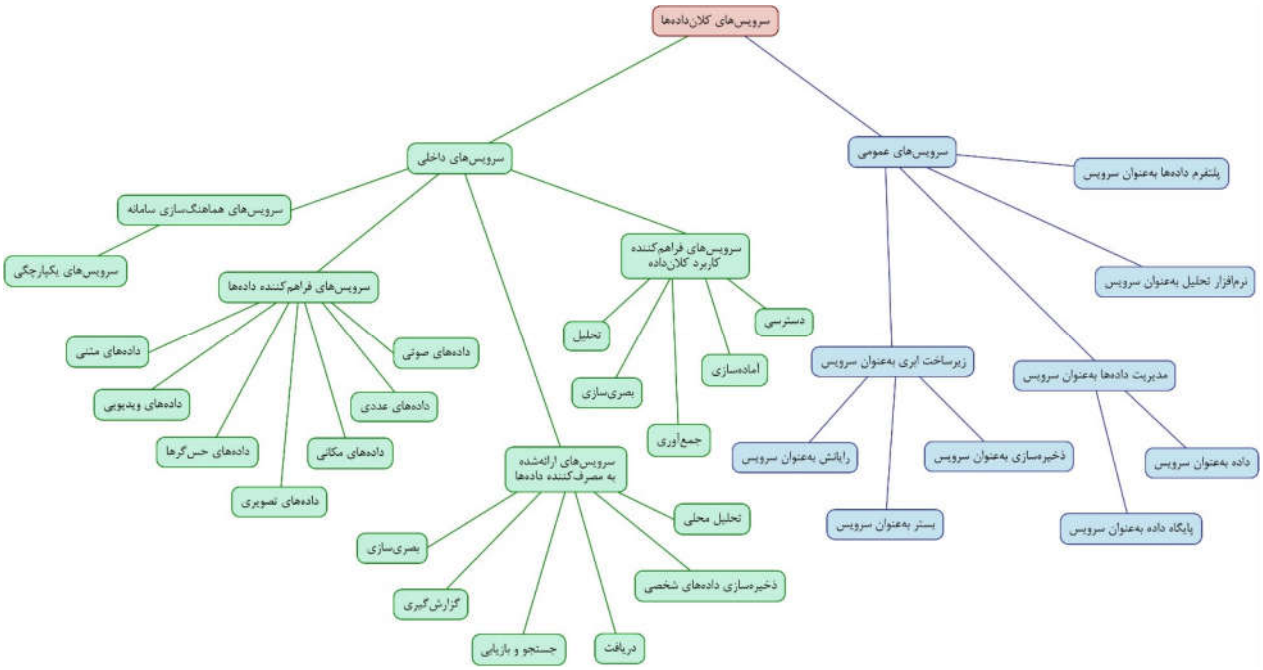
هدوپ: چارچوبی متن‌باز است که برای پردازش و ذخیره‌سازی کلان‌داده‌ها به‌کار می‌رود.

HDFS: سامانه فایل توزیع‌شده هدوپ است که برای ذخیره‌سازی داده‌ها به‌صورت توزیع‌شده و قابل تحلیل در چارچوب هدوپ به‌کار می‌رود.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۳ گونه‌شناسی

برای فراهم‌کنندگان سرویس، روش‌های مختلفی جهت ارائه سرویس در بازار کلان‌داده‌ها وجود دارد. این سرویس‌ها را می‌توان از منظر تجرید[۱] در لایه‌های مختلف دسته‌بندی کرد. در هرکدام از این لایه‌ها، خدمات متفاوتی قابل ارائه است که در شکل ۱ نشان داده شده است. لایه‌های مختلف سرویس‌های کلان‌داده‌ها که در این شکل نشان داده شده‌اند، نوع فناوری‌های استفاده‌شده در عملکرد آن دسته از سرویس‌ها را مشخص می‌کنند. برای مثال، لایه نرم‌افزار تحلیل به‌عنوان سرویس شامل فناوری‌هایی مانند Tibco Spot Silver است که یک پلتفرم تحلیل مبتنی بر محیط ابریست. یا سرویس S3 شرکت آمازون که سرویس‌های ذخیره‌سازی را فراهم می‌کند در لایه ذخیره‌سازی به‌عنوان سرویس قرار می‌گیرد. راحتی استفاده از سرویس در لایه‌های بالاتر، بیشتر می‌شود. هر لایه یک استفاده خاص داشته و سطح تجرید متفاوتی از پیچیدگی پردازش توزیع‌شده کلان‌داده‌ها را به کاربران نهایی ارائه می‌دهد.



شکل ۱- سرویس‌های کلان‌داده‌ها

[۱] Abstraction

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۳-۱ سرویس‌های داخلی

۳-۱-۱ سرویس‌های هماهنگ‌کننده سامانه

نقش هماهنگ‌کننده شامل تعریف و یکپارچه‌سازی فعالیت‌های کاربرد داده‌ها در یک سامانه عمودی عملیاتی است. معمولاً هماهنگ‌کننده سامانه درگیر مجموعه‌ای از نقش‌های خاص‌تر است که توسط یک یا چند بازیگر انجام می‌شوند و عملکرد سامانه کلان‌داده‌ها را مدیریت و هماهنگ می‌کنند. کارکرد هماهنگ‌کننده سامانه، پیکربندی و مدیریت سایر اجزای معماری کلان‌داده‌ها برای پیاده‌سازی یک یا چند بار کاری است که آن معماری برای اجرای آنها طراحی شده است. بارهای کاری مدیریت‌شده توسط هماهنگ‌کننده سامانه می‌تواند تخصیص اجزای چارچوب به تک‌تک گره‌های فیزیکی یا مجازی در سطح پایین‌تر یا فراهم کردن یک رابط کاربری گرافیکی پشتیبانی‌کننده از مشخصات روندهای کاری متصل‌کننده چند کاربرد و جزء به هم در سطح بالاتر باشد. همچنین هماهنگ‌کننده سامانه می‌تواند از طریق واحد مدیریت، بارهای کاری و سامانه را پایش کند تا دستیابی به نیازمندی‌های یک کیفیت سرویس خاص برای هر بار کاری را تأیید کند و به‌صورت پویا منابع فیزیکی و مجازی اضافه‌تر برای دستیابی به نیازمندی‌های بار کاری را تهیه کرده و اختصاص دهد.

در یک سامانه سازمانی، نقش هماهنگ‌کننده سامانه معمولاً متمرکز است و می‌تواند به نقش سنتی اداره‌کننده سامانه نگاشت شود که نیازمندی‌ها و محدودیت‌های سراسری که سامانه باید داشته باشد - شامل نیازمندی‌های سیاست، معماری، منابع یا کسب‌وکار - فراهم می‌کند. اداره‌کننده سامانه با مجموعه‌ای از سایر نقش‌ها (مانند مدیر داده‌ها، امنیت داده‌ها و مدیر سامانه) کار می‌کند تا نیازمندی‌ها و کارکردهای سامانه را پیاده‌سازی کند.

در یک سامانه عمودی، نقش هماهنگ‌کننده سامانه معمولاً غیرمتمرکز است. هر ذینفع مستقل، مسئول مدیریت، امنیت و یکپارچگی سامانه خود و همچنین یکپارچگی با سامانه توزیع‌شده کلان‌داده بااستفاده از سرویس‌ها و رابط‌های فراهم‌شده توسط سایر ذینفعان است.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۳-۱-۲ سرویس‌های فراهم‌کننده داده‌ها

فراهم‌کننده داده‌ها، مجموعه داده‌ها یا اطلاعات جدید را برای کشف، دسترسی و تبدیل توسط سامانه کلان‌داده‌ها، وارد این سامانه می‌کند. مجموعه داده‌های جدید با داده‌هایی که از قبل در سامانه استفاده شده‌اند و در مخازن مختلف قرار دارند، متفاوتند. فناوری‌های مشابهی می‌تواند برای دسترسی به هردو مجموعه داده جدید و موجود مورد استفاده قرار گیرد. کنش‌گران فراهم‌کننده داده می‌تواند هرچیزی از یک حس‌گر گرفته تا داده‌های واردشده به‌صورت دستی توسط انسان یا یک سامانه کلان‌داده‌های دیگر را شامل شود.

یکی از ویژگی‌های مهم یک سامانه کلان‌داده‌ها، توانایی آنها در دریافت و استفاده داده‌ها از منابع داده متنوع است. منابع داده می‌توانند مدارک، تصاویر، صدا، ویدیو، داده‌های حس‌گر، لاگ‌های وب، لاگ‌های سامانه، کوکی‌های HTML و سایر منابع عمومی یا داخلی باشند. انسان‌ها، ماشین‌ها، حس‌گرها، کاربردهای برخط و برون‌خط، فناوری‌های اینترنت و سایر کنش‌گران می‌توانند منابع داده را ایجاد کنند. نقش‌های فراهم‌کننده داده و فراهم‌کننده کاربرد کلان‌داده‌ها اغلب متعلق به سازمان‌های متفاوتی هستند مگر سازمانی که کاربرد کلان‌داده‌ها را توسعه می‌دهد، مالک منابع داده نیز باشد. بنابراین داده‌ها از منابع مختلف ممکن است ملاحظات امنیتی و حریم خصوصی متفاوتی داشته باشند. در مورد منابع داده‌های خام، فراهم‌کننده داده‌ها می‌توانند داده‌ها را پاک‌سازی، تصحیح و در یک فرمت داخلی که قابل دسترسی برای سامانه کلان‌دادهای مربوطه است، ذخیره کنند. فراهم‌کننده داده می‌تواند یک مرحله تجرید روی داده‌هایی که قبلا توسط سامانه دیگری تبدیل شده‌اند، انجام دهد. در این حالت، فراهم‌کننده داده مصرف‌کننده داده‌های یک سامانه کلان‌داده دیگر خواهد بود.

فراهم‌کننده داده، یک مجموعه از سرویس‌ها (واسط‌ها) را برای کشف و دسترسی داده‌ها ارائه می‌دهد. این سرویس‌ها معمولاً دارای یک ثبات هستند و بنابراین کاربردها می‌توانند یک فراهم‌کننده داده را مکان‌یابی کرده، داده‌های موردنظر که آن فراهم‌کننده دارد را شناسایی کرده، نوع دسترسی‌های مجاز را فهمیده، نوع تحلیل‌های پشتیبانی‌شده را فهمیده، منبع داده‌ها را مکان‌یابی کرده، روش‌های دسترسی را معین کرده، نیازمندی‌های امنیت داده‌ها را شناسایی کرده، نیازمندی‌ای حریم خصوصی داده‌ها را شناسایی کرده و سایر اطلاعات مربوطه را استخراج کنند. بنابراین سرویس

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

ارائه‌شده ابزارهای لازم برای ثبت منبع داده‌ها، جستار ثبات و شناسایی یک مجموعه داده استاندارد موجود در ثبات را فراهم می‌کند.

۳-۱-۳ سرویس‌های فراهم‌کننده کاربرد کلان‌داده‌ها

فراهم‌کننده کاربرد کلان‌داده‌ها، مجموعه عملیات خاصی را همراه با چرخه حیات داده‌ها اجرا می‌کند تا نیازمندی‌های اعلام‌شده توسط هماهنگ‌کننده سامانه و همچنین نیازمندی‌های امنیتی و حریم خصوصی را برآورده سازد. فراهم‌کننده کاربرد کلان‌داده‌ها، جزئی در معماری کلان‌داده‌هاست که منطق کسب‌وکار و عملکردی را که باید توسط معماری اجرا شود، کپسوله می‌کند. فعالیت‌های این نقش شامل موارد زیر است:

- جمع‌آوری
- آماده‌سازی
- تحلیل
- بصری‌سازی
- دسترسی

این فعالیت‌ها توسط زیراجزای فراهم‌کننده کاربرد کلان‌داده‌ها و در قالب رابط یا سرویس به سایر اجزای معماری مخصوصاً مصرف‌کننده داده‌ها ارائه می‌شود.

۳-۱-۴ سرویس‌های ارائه‌شده به مصرف‌کننده داده‌ها

مشابه فراهم‌کننده داده‌ها، مصرف‌کننده داده‌ها نیز می‌تواند کاربر نهایی یا یک سامانه دیگر باشد. به دلایل زیادی مصرف‌کننده داده‌ها تصویر آینه‌ای فراهم‌کننده داده‌هاست. فعالیت‌های مصرف‌کننده شامل موارد زیر هستند:

- جستجو و بازیابی
- دریافت
- تحلیل محلی

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

- گزارش‌گیری
- بصری‌سازی
- داده‌ها برای مصرف در پردازش‌های شخصی خودشان

مصرف‌کننده داده‌ها، رابط‌ها یا سرویس‌های فراهم‌شده توسط فراهم‌کننده کاربرد کلان‌داده‌ها را استفاده می‌کند تا به اطلاعات موردعلاقه خود دسترسی یابد. این سرویس‌ها و رابط‌ها می‌توانند دربرگیرنده گزارش‌گیری، بازیابی داده‌ها و نمایش داده‌ها باشند.

مصرف‌کننده داده‌ها معمولاً با فراهم‌کننده کاربرد کلان‌داده‌ها و از طریق رابط‌ها و سرویس‌های فوق تعامل می‌کند تا تحلیل و بصری‌سازی پیاده‌سازی‌شده توسط فراهم‌کننده کاربرد کلان‌داده‌ها استفاده کند. این تعامل ممکن است درخواست‌محور باشد که مصرف‌کننده داده‌ها دستور یا تراکنش را آغاز می‌کند و فراهم‌کننده کاربرد کلان‌داده‌ها به آن درخواست پاسخ می‌دهد. تعامل می‌تواند شامل بصری‌سازی‌های تعاملی، ایجاد گزارش‌ها یا کند و کاو در داده‌ها بااستفاده از عملکردهای هوشمندی کسب‌وکار فراهم‌شده توسط فراهم‌کننده کاربرد کلان‌داده‌ها باشد. یا اینکه تعامل می‌تواند رشته‌محور یا ارسال‌محور باشد که در این روش، مصرف‌کننده داده منتظر یک یا چند خروجی خودکار از کاربرد می‌ماند. تقریباً در تمامی حالت‌ها، واحد امنیت و حریم خصوصی معماری کلان‌داده‌ها، احراز هویت و احراز دسترسی بین مصرف‌کننده داده‌ها و معماری را پشتیبانی می‌کند. مشابه سرویس یا رابط بین معماری کلان‌داده‌ها و فراهم‌کننده داده‌ها، سرویس یا رابط بین مصرف‌کننده داده‌ها و فراهم‌کننده کاربرد کلان‌داده‌ها نیز سه فاز مجزای آغاز، انتقال داده‌ها و اختتام دارد.

۳-۲ سرویس‌های عمومی

۳-۲-۱ زیرساخت ابری به‌عنوان سرویس

هر زیرساخت کلان‌داده‌ها به‌عنوان سرویس، معمولاً دارای اجزای زیرساخت به‌عنوان سرویس - به‌خصوص رایانش به‌عنوان سرویس (CaaS) و ذخیره‌سازی به‌عنوان سرویس - است. همچنین بخش زیادی از کلان‌داده‌ها معمولاً در

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

نرم‌افزارهای کاربردی زیرساخت ابری فراهم‌کننده سرویس مورد استفاده قرار می‌گیرند. انتقال حجم زیادی از داده‌ها می‌تواند در برخی سناریوها محدودکننده باشد. از این‌رو داشتن داده‌ها (که باید بیشتر پردازش شوند) در زیرساخت فراهم‌کننده، فراهم‌کنندگان سرویس‌های کلان‌داده‌ها را قادر به بهینه‌سازی سرویس خود برای سرویس‌های ارائه‌شده از همان زیرساخت می‌کند.

در این لایه، پلتفرم‌های ابری مانند open stack و سرور VMware ESX محیط ابری مجازی را فراهم می‌کنند که مبنای سرویس‌های کلان‌داده‌ها را شکل می‌دهد. همچنین این لایه پایش استفاده[۲] و صورتحساب برای سرویس‌ها را نیز فراهم می‌کند. گرچه این لایه بخشی از سایر انواع سرویس‌های کلان‌داده‌ها نیز به‌شمار می‌آید ولی از طریق نرم‌افزارهای کاربردی خارجی قابل دسترسی نیست.

۳-۲-۱-۱ *رایانش به‌عنوان سرویس*

این سرویس‌ها از فناوری‌هایی تشکیل شده‌اند که سرویس‌های رایانش را روی یک بستر وب فراهم می‌کنند. برای مثال با استفاده از EMR[۳] کاربران می‌توانند برنامه‌هایی را برای دستکاری داده‌ها و ذخیره‌سازی نتایج در یک بستر ابری اجرا کنند. این لایه در کنار چارچوب‌های پردازشی، شامل APIها و دیگر برنامه‌های کمکی نیز می‌شود.

۳-۲-۱-۲ *ذخیره‌سازی به‌عنوان سرویس*

این لایه معمولاً در یک سیستم فایل پرتابل مقیاس‌پذیر، توزیع شده است. این سیستم فایل می‌تواند سیستم فایل HDFS ساخته‌شده با نام گره و خوشه‌ای از گره‌های داده‌ها باشد که بنا به درخواست ظرفیت آن می‌تواند توسعه یابد (مقیاس‌پذیری). سرویس‌های داده به‌عنوان سرویس معمولاً دارای یک جزء ارائه هستند که کاربران با استفاده از آن می‌توانند به‌طور مستقیم با سیستم فایل HDFS تعامل داشته و داده‌های خود را به‌منظور تحلیل روی آن قرار دهند.

---

[۲] Usage monitoring

[۳] Amazon Elastic MapReduce

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۳-۲-۲مدیریت داده‌ها به‌عنوان سرویس

لایه دیگری از سرویس‌های کلان‌داده‌ها، سرویس‌های مدیریت داده‌هاست. سرویس‌هایی مانند پایگاه داده به‌عنوان سرویس (DBaaS) یا تجمیع داده‌ها و ارائه یکپارچه آنها که داده به‌عنوان سرویس (DaaS) شناخته می‌شود، در این لایه قرار می‌گیرند.

در این لایه، نرم‌افزارهای کاربردی سطح بالاتری مانند سرویس پایگاه داده رابطه‌ای (RDS) آمازون و DynamoDB پیاده‌سازی شده‌اند تا سرویس‌های مدیریت و پردازش توزیع‌شده داده را فراهم کنند. فناوری‌های موجود در این لایه‌ها سرویس‌های مدیریت پایگاه داده را روی بستر ابری فراهم می‌کنند. برای مثال، اگر کاربری نیاز به یک پایگاه داده Oracle در محیط ابری داشته باشد، با استفاده از RDS آمازون می‌تواند چنین سرویسی را در اختیار داشته باشد. از آنجاکه سرویس‌های مدیریت داده‌ها توسط فراهم‌کننده سرویس مدیریت می‌شوند (مانند گرفتن پیشتیبان از داده‌ها)، این سرویس‌ها استفاده را آسان‌تر کرده و نیازمندی‌های منابع را برای نگه‌داری داده‌ها کاهش می‌دهند.

۳-۲-۲-۱ داده به‌عنوان سرویس

داده به‌عنوان سرویس اصولاً اشاره به تجمیع و مدیریت یک مجموعه داده خاص و اجازه دسترسی کنترل‌شده به آن مجموعه داده از طریق یک API دارد. یک مثال از این سرویس، سرویس داده‌های عمومی گوگل است که دسترسی به همه نوع داده فراهم‌شده توسط پژوهشگاه‌های عمومی امریکا را میسر می‌سازد که می‌توان از طریق این سرویس از آن داده‌ها در نرم‌افزارهای کاربردی برای بصری‌سازی استفاده کرد. مثال دیگر، مخزن داده‌های آکادمیک و بیوانفورماتیک ملی امریکاست که شرکت‌های دارویی می‌توانند تحلیل پژوهش‌های خود را روی دادگان آن انجام دهند. فراهم‌کننده سرویس در چنین سناریویی، نقش تجمیع‌کننده و تصدی‌کننده داده‌ها را دارد.

۳-۲-۲-۲ پایگاه داده به‌عنوان سرویس

نوع دیگری از سرویس‌های مدیریت داده‌ها، سرویس‌های زیرساختی مدیریت داده‌هاست که به آن پایگاه داده به‌عنوان سرویس (DBaaS) گفته می‌شود. اینگونه سرویس‌های پایگاه داده در دسترس نرم‌افزارهای کاربردی به‌کار

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

گرفته‌شده در هر محیط اجرایی قرار می‌گیرند که شامل پلتفرم به‌عنوان سرویس (PaaS) نیز می‌شود. این پایگاه‌های داده می‌توانند پایگاه‌های داده رابطه‌ای سنتی (SQL) نیز باشند ولی در حوزه کلان‌داده‌ها معمولاً پایگاه‌های داده NoSQL و درون‌حافظه‌ای[۴] مورد استفاده قرار می‌گیرند.

۳-۲-۳ پلتفرم داده‌ها به‌عنوان سرویس

در این لایه از سرویس‌ها، فراهم‌کننده سرویس نه‌تنها یک زیرساخت مدیریت داده را ارائه می‌دهد بلکه محیط اجرا برای نرم‌افزارهای کاربردی و اسکریپت‌های پردازش داده‌ها را نیز در سرویس خود فراهم می‌کند. بنابراین، کاربران می‌توانند هم داده‌ها و هم کدهای تحلیل خود را ارسال کنند و پلتفرم، خوشه‌بندی داده‌ها، ایجاد و اجرای گره‌های پردازشی را به‌صورت خودکار انجام می‌دهد. البته کاربران در این سناریو افرادی فنی شامل دانشمندان داده و برنامه‌نویسان هستند که می‌توانند مدیریت محیط‌های تحلیل اختصاصی را همراه با نوشتن کدهای تحلیل داده‌ها، انجام دهند.

۳-۲-۴نرم‌افزار تحلیل به‌عنوان سرویس

در مقابل پلتفرم داده‌ها به‌عنوان سرویس، کاربران نرم‌افزار تحلیل به‌عنوان سرویس با تعامل با یک پلتفرم تحلیل در یک سطح تجرید بالاتر، بیشتر آشنا هستند زیرا معمولاً یا باید اسکریپت‌ها و جستارهایی[۵] را اجرا کنند که دانشمندان داده‌ها یا برنامه‌نویسان برای آنها توسعه داده‌اند یا گزارش‌ها، بصری‌سازی‌ها و داشبوردها را آماده کنند.

نرم‌افزار تحلیل به‌عنوان سرویس معمولاً وابسته به محصول یک شرکت نرم‌افزاری و بازار هدف خاص‌منظوره آن است. در حالی‌که سرویس‌های مدیریت داده‌ها و پلتفرم داده‌ها به‌عنوان سرویس معمولاً تمایل به ارائه راهکارهای عام‌منظوره دارند، نرم‌افزار تحلیل به‌عنوان سرویس بیشتر راهکارهای خاص‌منظوره را ارائه می‌دهد. بنابراین

[۴] In-memory

[۵] Queries

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

فراهم‌کنندگان سرویسی که می‌خواهند وارد بازار نرم‌افزار تحلیل به‌عنوان سرویس شوند باید انتخاب کنند که قصد ارائه سرویس به کدام حوزه‌ها و صنایع را دارند.

این لایه شامل نرم‌افزارهای کاربردی سطح بالا مشابه R و Tableau است که روی یک پلتفرم رایانش ابری، ارائه شده و برای تحلیل داده‌ها استفاده می‌شوند. کاربران می‌توانند از طریق یک واسط وب به این سرویس‌ها دسترسی داشته باشند که از آنجا می‌توانند جستارها را ایجاد کرده و گزارش‌های مبتنی بر داده‌های موجود در لایه ذخیره‌سازی را تعریف کنند. فناوری‌های استفاده‌شده در لایه تحلیل داده‌ها، پیچیدگی‌های فنی BDaaS را مخفی کرده و استفاده بهتر از داده‌ها در سامانه را ممکن می‌سازند. واسط وب این فناوری‌ها ممکن است دارای ابزارهای گرافیکی برای انجام تحلیل‌های آماری پیچیده توسط کاربران باشد. برای مثال اگر داده‌های موجود در یک فایل متنی باید تحلیل شوند، کاربر می‌تواند آن فایل را به پلتفرم ابری Tableau ارسال کرده و سپس تحلیل را با انجام چند مرحله تنظیمات ساده به‌صورت Wizard، انجام دهد. فناوری‌ها در این لایه می‌توانند بسته به صنعت، خاص‌منظوره‌تر باشند. برای مثال، یک فراهم‌کننده نرم‌افزار تحلیل به‌عنوان سرویس می‌تواند نرم‌افزاهای تحلیل تخصصی برای سرویس‌های مالی با توان تحلیل و بصری‌سازی‌های معمول در این صنعت را ارائه دهد. این سرویس شامل عملکردهای پایش مخاطرات، انجام امور بانکی و شبیه‌سازی قیمت سهام می‌شود.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۴ سرویس‌های موفق کلان‌داده‌ها در جهان

بازار سرویس‌ها و راهکارهای مبتنی بر کلان‌داده‌ها به‌صورت مدام در حال رشد است که نشان‌دهنده رشد تعداد شرکت‌های فراهم‌کننده آنهاست. در مطالعه‌ای که گروه Experton به سفارش T-Systems در سال ۲۰۱۶ انجام داده است، بیش از ۲۰۰ شرکت استفاده‌کننده از کلان‌داده‌ها شناسایی شده که در مقایسه با ۱۳۸ شرکت شناسایی‌شده در سال ۲۰۱۵ رشد چشم‌گیری داشته است[۵]. از آنجا که مجال بررسی تمامی این شرکت‌ها در این گزارش وجود ندارد، لذا فقط به بررسی موفق‌ترین این شرکت‌ها و نیز سرویس‌های ارائه‌شده توسط آنها اکتفا می‌کنیم.

نکته دیگری که وجود دارد این است که سرویس‌های داخلی کلان‌داده‌ها که در بخش قبلی توضیح داده شد، معمولاً سرویس‌هایی هستند که در قالب یک قرارداد بین دو شرکت که بازیگران اکوسیستم کلان‌داده‌ها هستند، ارائه می‌شوند و بنابراین اطلاعات چندانی در مورد این سرویس‌ها در دسترس نیست. این دسته از سرویس‌ها بنا به نیاز شرکت متقاضی طراحی و ارائه می‌شوند و به‌همین جهت نمی‌توان یک چارچوب یا استاندارد کلی برای آنها معین کرد. برخی از این سرویس‌ها نیز توسط فراهم‌کننده سرویس، بااستفاده از سرویس‌های عمومی کلان‌داده‌ها یا ترکیبی از آنها ارائه می‌شوند که سرویس‌های شاخص در این حوزه را در این بخش توضیح می‌دهیم.

۴-۱ آمازون

شرکت آمازون سرویس‌های متنوعی را در حوزه کلان‌داده‌ها ارائه می‌دهد که هرکدام از آنها را به‌صورت جداگانه مورد بررسی قرار می‌دهیم.

۴-۱-۱ EC2

EC2، سرویس پایه آمازون در رایانش ابری است که در سال ۲۰۰۶ معرفی شد. این سرویس یک IaaS است که کاربران را قادر به دسترسی به رایانش ابری روی پلتفرم ابری آمازون می‌کند. با استفاده از IaaS، کاربران EC2 می‌توانند منابع مجازی را روی پلتفرم رایانش ابری ایجاد کنند و سپس نرم‌افزارهای کاربردی خود را روی آن توسعه دهند. با اینکه این سرویس، ذاتاً یک سرویس کلان‌داده محسوب نمی‌شود اما با توجه به ساختار و امکانات آن، در پیاده‌سازی

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

کاربردهای کلان‌داده بسیار پرکاربرد است و به همین دلیل در این قسمت به شرح آن می‌پردازیم. امکانات و ساختار این سرویس که باعث مناسب بودن آن برای توسعه کاربردهای کلان‌داده شده نیز در ادامه توضیح داده می‌شود.

EC2 اجازه می‌دهد که کاربران، ماشین‌های رایانش مجازی خود را داشته باشند که مطابق با نیازهای آنها ساخته شده‌اند. همچنین آمازون به کاربران خود این امکان را می‌دهد که موقعیت جغرافیایی سرورهای مجازی خود را انتخاب کنند. به این ترتیب، کاربران می‌توانند با انتخاب سرورهای مجازی در مناطق جغرافیایی مختلف که مراکز داده آمازون قرار دارند، پلتفرم رایانشی خود را به‌صورت توزیع‌شده، ایجاد کنند. این قابلیت از یک جهت دیگر هم برای کاربران سودمند است که آن، مقررات رگولاتوری داده‌هاست. برای مثال سازمان‌های وابسته به دولت ایالات متحده مجاز به نگه‌داری داده‌های خود در خارج از مرزهای این کشور نیستند. کشورهای اروپایی نیز قوانین سخت‌گیرانه‌ای در مورد حفاظت از داده‌ها مخصوصاً داده‌های شخصی مربوط به شهروندان اروپایی دارند.

از آنجا که سرویس EC2 آمازون یک سامانه IaaS بدون هیچ‌گونه قابلیت مختص کلان‌داده‌ها است، آن را می‌توان در لایه‌های پایین گونه‌شناسی سرویس‌های کلان‌داده‌ها قرار داد. هیچ قابلیت ذاتی کلان‌داده‌ها و تحلیل داده‌ها در EC2 وجود ندارد و بنابراین کاربران باید خودشان این قابلیت‌ها را پیاده‌سازی کنند. کاربران باید بعد از ایجاد ماشین‌های مجازی، نرم‌افزارهای رایانش توزیع‌شده مانند HDFS را نصب کنند تا بتوانند سامانه‌های کلان‌داده‌ها را پیاده‌سازی کنند. یک سرویس IaaS پایه به کاربران اجازه سفارشی‌سازی سامانه‌های خود را می‌دهد اما برای این منظور نیاز به تخصص‌ها و مهارت‌های خاصی است. برای استفاده از EC2 در کلان‌داده‌ها، کاربران باید ماشین‌های مجازی را ایجاد کرده و سپس معماری لازم را نصب کنند. برای مثال کاربران می‌توانند یک سامانه کلان‌داده‌ها توزیع‌شده مانند Hadoop را با ایجاد الگوهای AMI برای سرورها روی یک سامانه EC2 آمازون به‌وجود آورند و در مواقع نیاز، گره‌های پردازشی جدید را به سامانه خود بیافزایند. بنابراین زمانی‌که کاربر قصد سامانه کلان‌داده‌ها مبتنی بر EC2 خود را گسترش دهد، می‌تواند گره‌های داده از پیش تنظیم‌شده توسط الگوهای AMI را اضافه کند. EC2 قابلیت‌های ابری لازم موردنیاز برای ایجاد سامانه‌های کلان‌داده‌ها را فراهم می‌کند.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

Dynamo ۲-۱-۴

سرویس Dynamo آمازون یک سرویس پایگاه داده NoSQL مبتنی بر محیط ابریست که کاربران را قادر به ذخیره‌سازی و بازیابی داده‌ها در حجم‌های زیاد می‌کند. پایگاه‌های داده NoSQL محدودیت‌های کمتری روی مدل‌های پایداری[۶] (مشخصه‌های ACID) نسبت به پایگاه‌های داده رابطه‌ای مانند Oracle دارند. پایگاه‌های داده NoSQL معمولاً بخشی از فناوری‌های کلان‌داده‌ها فرض می‌شود که دلیل آن، ظرفیت آنها در ذخیره‌سازی دادگان بزرگ و ساختارنیافته است.

DynamoDB آمازون مبتنی بر چارچوب Dynamo برای ذخیره‌سازی کلید-مقدار[۷] توزیع‌شده است. سرویس DynamoDB آمازون با سایر سرویس‌های این شرکت متفاوت است زیر کاربران آن برپایه توان عملیاتی شارژ می‌شوند نه فضای ذخیره‌سازی. بنابراین هنگام گسترش سرویس، کاربر کافیست درخواست ارائه توان عملیاتی بالاتر از طرف سرویس DynamoDB دهد.

DynamoDB یک سرویس ذخیره‌سازی مقیاس‌پذیر ارائه‌شده توسط آمازون است و سه لایه سرویس ذخیره‌سازی به‌عنوان سرویس، رایانش به‌عنوان سرویس و مدیریت داده‌ها به‌عنوان سرویس را فراهم می‌کند. این سرویس، پیچیدگی ذخیره کننده داده NoSQL را تجرید می‌کند و کاربران را قادر می‌سازد تا از طریق یک واسط ارائه، به داده‌ها دسترسی داشته باشند. این پایگاه داده مشابه سایر پایگاه‌های داده کلید-مقدار مانند Cassandra و MongoDB است با این تفاوت که به صورت سرویس توسط آمازون ارائه می‌شود.

Elastic MapReduce ۳-۱-۴

سرویس EMR آمازون سرویسی است که چارچوب پردازش توزیع‌شده Hadoop را برای پردازش مؤثر مجموعه داده‌های بزرگ مورد استفاده قرار می‌دهد. EMR کاربران را قادر می‌سازد که از زیرساخت Hadoop به‌صورت کاربرپسند و بدون نیاز به دانستن نحوه نصب یا مدیریت زیرساخت، استفاده کنند. از آنجا که EMR از یک چارچوب

[۶] Consistency

[۷] Key-value

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

توزیع‌شده استفاده می‌کند، با سرعت و به‌طور مؤثر داده‌های بسیار زیادی را روی یک خوشه مقیاس‌پذیر EC2 آمازون از طریق مقیاس‌بندی افقی، پردازش می‌کند. این سرویس در حوزه‌های مختلفی مانند تحلیل رویدادها، انبار کردن داده‌ها[۸] و یادگیری ماشینی مورد استفاده قرار می‌گیرد. ایجاد خوشه Hadoop و مدیریت کارهای MapReduce به‌همراه انتقال داده‌ها در EMR به‌صورت خودکار صورت می‌پذیرد. EMR لایه‌های رایانش به‌عنوان سرویس، ذخیره‌سازی به‌عنوان سرویس و زیرساخت ابری به‌عنوان سرویس را پوشش می‌دهد.

۴-۲ BigQuery گوگل

BigQuery گوگل کاربران را قادر به دسترسی آنی به مجموعه داده‌های وسیع بدون نیاز به ایجاد زیرساخت گران‌قیمت می‌کند[۶]. با استفاده از این سرویس کاربران می‌توانند در مصرف منابع صرفه‌جویی کنند در حالی‌که آنها را توانمند می‌سازد تا با یک پلتفرم ابری کاربرپسند، جستارهای خود را روی دادگان خود انجام دهند. BigQuery گوگل از طریق یک واسط کاربری مبتنی بر وب برای انجام جستارها قابل دسترسی است.

BigQuery گوگل بر پایه Dewmel است که یک سامانه جستار مقیاس‌پذیر Ad-hoc برای کاوش دادگان بزرگ است. Dremel از درخت‌های چندسطحی استفاده می‌کند و دارای چینش ستونی است. این سامانه یک سرویس جستار است که کاربران می‌توانند برای اجرای جستارهای شبیه به SQL به‌هدف اجرای جستارهای تجمیعی سریع روی دادگانی با بیش از یک تریلیون سطر استفاده کنند. Dremel و MapReduce دو فناوری مکمل هستند که جای همدیگر را پر نمی‌کنند.

از آنجا که سرویس BigQuery گوگل سرویس‌های رایانش و تحلیل داده‌ها را ارائه می‌دهد، می‌توان ذخیره‌سازی داده‌ها برای سرویس BigQuery گوگل توسط منبع ذخیره Cloud Storage انجام می‌شود که سرویسی مشابه سرویس S3 آمازون است. هر دو سرویس، ذخیره‌سازی انبوه داده‌ها در پلتفرم ابری را فراهم می‌کنند. سرویس BigQuery

[۸] Data warehousing

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

گوگل دارای قابلیت‌های نمایش و تحلیل داده‌ها نیست و از این‌رو آن را نمی‌توان همچون سرویس AWS یک سرویس تحلیل داده‌ها به‌عنوان سرویس در نظر گرفت.

۴-۳ Splunk Storm

Splunk Storm یک پلتفرم تحلیل داده‌ها برای پایش و تحلیل داده‌های لاگ‌هاست. این سرویس کاربران را قادر می‌سازد تا یک مخزن داده ایجاد کنند که آسان‌تر کردن تولید گراف‌های، داشبوردها و سایر واسط‌های بصری که اطلاعات را به‌طور آنی فراهم می‌کنند، نمایه‌گذاری می‌شود[۷]. Splunk الگوهای داده‌ها را شناسایی کرده و کمک می‌کند تا با استفاده از یک واسط مبتنی بر وب، اطلاعات در درون سازمان منتشر شود.

Splunk از یک پایگاه داده اختصاصی به نام «index» برای ذخیره داده‌های کاربران استفاده می‌کند. این داده‌ها به‌صورت درونی در دو نوع دسته‌بندی می‌شوند. دسته اول، داده‌های خام هستند و دسته دوم، نمایه‌هایی هستند که به داده‌های خام ذخیره‌شده در پلتفرم ابری اشاره می‌کنند. محیط ابری Splunk روی یک پلتفرم AWS آمازون میزبانی[۹] شده است و روی ماشین‌های لینوکس ۶۴ بیت اجرا می‌شود. این سرویس، لایه‌های نرم‌افزار تحلیل به‌عنوان سرویس، رایانش به‌عنوان سرویس، مدیریت داده‌ها به‌عنوان سرویس و ذخیره‌سازی داده‌ها به‌عنوان سرویس را فراهم می‌کند ولی زیرساخت ابری آن سرویس‌های AWS آمازون هستند.

۴-۴ Azure-HDInsight مایکروسافت

HDInsight یک سرویس کلان‌داده‌ها ارائه‌شده توسط مایکروسافت است که اجازه ایجاد خوشه‌های Hadoop در یک محیط ابری را می‌دهد[۸]. این سرویس، Apache Hadoop را به‌عنوان سرویس فراهم می‌کند که کاربران نهایی را قادر به داشتن یک محیط مقیاس‌پذیر و مقرون‌به‌صرفه می‌کند. HDInsight یک محصول پرچم‌دار در تحلیل کلان‌داده‌ها مبتنی بر محیط ابری شناخته می‌شود.

[۹] Host

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

HDInsight، پلتفرم ابری مایکروسافت را با معماری پردازش توزیع‌شده Hadoop ترکیب می‌کند تا به مشتریان روی یک زیرساخت درخواستی[۱۰] برای تحلیل دادگان بزرگ سرویس‌دهی کند. پلتفرم ابری مایکروسافت یک سرویس رایانش ابری مشابه AWS آمازون است. سرویس‌های ابری آمازون که قابلیت استفاده در کلان‌داده‌ها را دارند، نسبت به HDInsight بهتر با یکدیگر یکپارچه می‌شوند. از آنجا که HDInsight دارای سرویس ذخیره‌سازی پایه روی زیرساخت ابری IaaS است، لایه‌های زیرساخت ذخیره‌سازی به‌عنوان سرویس، رایانش به‌عنوان سرویس و مدیریت داده‌ها به‌عنوان سرویس را پوشش می‌دهد.

۴-۵ Tibco Silver Spotfire

Tibco Spotfire یک پلتفرم تحلیل داده است که کاربران را قادر به انجام تحلیل‌های پیچیده و تولید گزارش‌های تعاملی می‌کند[۹], [۱۰]. کاربران می‌توانند گزارش‌های تحلیلی پویا را تولید کنند که می‌توان از طریق وب روی پلتفرم تحلیل داده Spotfire به آنها دسترسی داشت. این سرویس قابلیت یکپارچه شدن با سایر محصولات شرکت Tibco را نیز داراست.

از آنجا که Silver Spotfire یک پلتفرم تحلیل مبتنی بر وب است، لایه نرم‌افزار تحلیل به‌عنوان سرویس را پوشش می‌دهد. از آنجا که این سرویس، داده‌های خامی را که باید تحلیل شوند روی ماشین کاربر نگه‌داری می‌کند و فقط نتایج تحلیل را به پلتفرم ابری خود انتقال می‌دهد، لایه مدیریت داده‌ها به‌عنوان سرویس را پوشش نمی‌دهد. همچنین در این سرویس، رایانش نیز در محیط ابری انجام نمی‌شود و بنابراین لایه رایانش به‌عنوان سرویس نیز توسط این سرویس فراهم نمی‌شود. در مقایسه با سایر سرویس‌های کلان‌داده‌ها مانند EMR آمازون یا BigQuery گوگل، این سرویس دارای واسط کاربری بهتری است.

[۱۰] On-demand

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

QuBole ۴-۶

QuBole یک شرکت فراهم‌کننده سرویس‌های کلان‌داده‌ها است که چارچوب Hadoop را روی پلتفرم مبتنی بر محیط ابری ارائه می‌دهد. این سرویس دارای قابلیت‌هایی مانند مقیاس‌بندی خودکار و واسط‌های کاربری گرافیکی برای مدیریت پیاده‌سازی کلان‌داده‌ها روی یک پلتفرم ابریست[۱۱]. QuBole پیچیدگی‌های مدیریت خوشه‌های Hadoop را مخفی کرده و مقیاس‌پذیری آنی را فراهم می‌کند. این سرویس به کاربران امکان می‌دهد تا مستقیماً به دادگان ذخیره‌شده روی یک پلتفرم ابری مانند S2 آمازون متصل شوند.

سرویس QuBole یک موتور Hadoop داخلی دارد که تخصیص منابع را برای اجرای سریع‌تر، بهینه می‌کند. کاربران می‌توانند داده‌ها را مستقیماً از سرویس S3 آمازون خود منتقل کرده و با استفاده از واسط کاربری گرافیکی، جدول‌های Hive را ایجاد کنند. خوشه‌های Hadoop روی حساب کاربری AWS مشتری ایجاد می‌شوند. هنگام ایجاد حساب کاربری QuBole، کاربران باید حساب‌های کاربری AWS خود را ثبت کنند که QuBole از این اطلاعات برای ایجاد خوشه‌ها استفاده می‌کند.

پلتفرم QuBole اصولاً سرویس‌های پردازش داده‌ها را با استفاده از محصولات Apache فراهم می‌کند و لایه‌های رایانش به‌عنوان سرویس و مدیریت داده‌ها به‌عنوان سرویس را پوشش می‌دهد. از آنجا که این سرویس از سرویس S3 آمازون برای ذخیره‌سازی داده‌ها بهره می‌گیرد، لایه ذخیره‌سازی به‌عنوان سرویس را فراهم نمی‌کند و از آنجا که فاقد پلتفرم تحلیل داده‌هاست، لایه نرم‌افزار تحلیل به‌عنوان سرویس را هم ارائه نمی‌دهد.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

۵ جمع‌بندی

در این گزارش، گونه‌شناسی سرویس‌های کلان‌داده‌ها توضیح داده شد. به‌طور کلی سرویس‌های کلان‌داده‌ها را می‌توان به دو دسته عمده تقسیم‌بندی کرد. دسته اول سرویس‌هایی هستند که در معماری سامانه‌های کلان‌داده‌ها، به بازیگران مختلف، توسط سایر بازیگران اکوسیستم ارائه می‌شوند. این دسته از سرویس‌ها، معمولاً به‌صورت خاص‌منظوره و براساس نیازمندی‌های اعلام‌شده توسط بازیگر متقاضی طراحی و پیاده‌سازی می‌شوند. در برخی از موارد، از سرویس‌های عمومی نیز در پیاده‌سازی سرویس‌های داخلی استفاده می‌شود.

دسته دوم سرویس‌های کلان‌داده‌ها، سرویس‌های عمومی هستند. این سرویس‌ها را می‌توان سرویس‌های سطح بالاتر سرویس‌های ابری دانست که لایه‌های تجریدی مختلفی را بر سرویس‌های ابری پایه اضافه می‌کنند. هرچه لایه تجرید در این سرویس‌های بالاتر برود، کاربرد آنها محدودتر شده و انعطاف‌پذیری کمتری خواهند داشت اما معمولاً این سرویس‌ها امکان دسترسی به لایه‌های تجریدی پایین‌تر را نیز به کاربران می‌دهند تا بااستفاده از این امکان، انعطاف‌پذیری خود را تا حدودی افزایش دهند.

مطالعه سرویس‌های موفق در این حوزه نشان می‌دهد که سرویس‌های عمومی کلان‌داده‌ها در سایر کشورها تا حدی به‌صورت تجاری ارائه شده‌اند که اکثر شرکت‌های ارائه‌دهنده این خدمات، شرکت‌های امریکایی هستند. **Error! Reference source not found.** مقایسه بین سرویس‌های بررسی‌شده را نشان می‌دهد. با اینکه چنین خدماتی ارائه شده است اما به‌نظر می‌رسد بسیاری از متقاضیان سرویس‌های کلان‌داده‌ها از سرویس‌های ابری پایه برای پیاده‌سازی معماری خود بهره می‌گیرند و سرویس‌های تخصصی‌تر کلان‌داده‌ها هنوز در ابتدای راه هستند. با وجود این و با توجه به افزایش تقاضای تحلیل کلان‌داده‌ها انتظار می‌رود این سرویس‌ها نیز به‌مرور زمان دوره تکامل خود را گذرانده و توسط متقاضیان، به‌صورت گسترده‌تری مورداستفاده قرار گیرند.

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

مقایسه بین سرویس‌های موفق بررسی‌شده

نام سرویس	نام شرکت ارائه‌دهنده	نوع سرویس	فناوری استفاده‌شده
EC2	آمازون	زیرساخت ابری به‌عنوان سرویس	مجازی‌سازی
Dynamo	آمازون	پایگاه داده به‌عنوان سرویس	پایگاه داده توزیع‌شده NoSQL
EMR	آمازون	رایانش به‌عنوان سرویس	Hadoop
BigQuery	گوگل	رایانش به‌عنوان سرویس	ذخیره‌سازی توزیع‌شده داده‌ها
Splunk Storm	Splunk	نرم‌افزار تحلیل به‌عنوان سرویس	تحلیل داده‌های لاگ
Azure-HDInsight	مایکروسافت	رایانش به‌عنوان سرویس	Hadoop
Tibco Silver Spotfire	Tibco	نرم‌افزار تحلیل به‌عنوان سرویس	تحلیل داده‌ها و تولید گزارش‌های تعاملی
QuBole	Qubole	رایانش به‌عنوان سرویس	Hadoop

گزارش: "سرویس‌های کلان‌داده‌ها	وضعیت: نهایی	کد

مراجع

[1] Jim Metzler and Steve Taylor, “Defining applications vs services,” *Network World*, 15-Aug-2011. [Online]. Available: http://www.networkworld.com/article/2180125/lan-wan/defining-applications-vs-services.html.
[2] “Computing platform,” *Wikipedia*. 27-Feb-2017.
[3] Benoy Bhagattjee, “Emergence and Taxonomy of Big Data as a Service.” May-2014.
[4] Mark A. Beyer, John-David Lovelock, Dan Sommer, and Merv Adrian, “Big Data Drives Rapid Changes in Infrastructure and $232 Billion in IT Spending Through 2016.” Gartner Inc., Oct-2012.
[5] Holm Landrock, Oliver Schonschek, Prof. Dr. Andreas Gadatsch, “Big Data Vendor Benchmark 2015.” Experton Group AG.
[6] “BigQuery - Google CloudPlatform.” [Online]. Available: https://cloud.google.com/bigquery/.
[7] “Splunk cloud - Product brief.” Splunk, 2016.
[8] Avkash Chauhan, Valentine Fontama, Michele Hart, Wee Hyong Tok, and Buck Woody, “Introducing Microsoft Azure HDInsight - Technical Overview.” Microsoft Press, 2014.
[9] “TIBCO Spotfire Platform.” [Online]. Available: http://spotfire.tibco.com/products/spotfire-platform.
[10] “TIBCO Spotfire Cloud.” [Online]. Available: spotfire.tibco.com/products/spotfire-cloud.
[11] “Qubole - Big Data platform tools and benefits.” [Online]. Available: https://www.qubole.com/features/.