به نام خدا

ارزیابی کردن پایگاه داده Cassandra NoSQL

درس: اصول طراحی پایگاه داده‌ها

استاد: طلعتیان آزاد

دانشجو: علی دشتی

شماره دانشجویی: 970216624

1- فناوری های پیشرفته سخت افزاری و نرم افزاری سرعت و کارایی انجام گردش کار علمی را افزایش می دهند. دانشمندان با مقایسه نتایج حاصل از این عمل ها و دقت بیشتر در تجزیه و تحلیل داده ها ، ممکن است یک گردش کار مشخص را بارها اجرا کنند. با این حال ، دست زدن به حجم زیادی از داده های تولید شده توسط اجرای برنامه های متمایز تحت شرایط مختلف به طور افزایشی دشوار می شود. این مقادیر عظیم داده ها باید ذخیره و درمان شوند تا از تحقیقات ژنومیک فعلی پشتیبانی کنند. بنابراین ، یکی از مشکلات اصلی هنگام کار با داده های ژنومیک ، ذخیره و جستجوی این داده ها است که به منابع محاسباتی زیادی نیاز دارد. در محیط های محاسباتی با مقادیر زیادی از داده های احتمالاً غیر متعارف ، سیستم های پایگاه داده NoSQL به عنوان جایگزینی برای سیستم های مدیریت پایگاه داده رابطه ای (RDBMS) سنتی ظاهر شده اند. سیستم های NoSQL پایگاه های داده توزیع شده ای هستند که برای پاسخگویی به خواسته های مقیاس پذیری بالا و تحمل خطا در مدیریت و تجزیه و تحلیل مقادیر داده ها بزرگ ساخته شده اند. پایگاه های داده NoSQL در بسیاری از زبان های برنامه نویسی مجزا کدگذاری شده اند و به طور کلی به عنوان نرم افزار منبع باز در دسترس هستند. هدف این مقاله بررسی تداوم داده های ژنومی بر روی یک سیستم پایگاه داده NoSQL خاص و پرکاربرد ، یعنی Cassandra است. آزمایشات انجام شده برای این مطالعه از داده های ژنومیک واقعی برای ارزیابی عملیات درج و استخراج به داخل و از پایگاه داده Cassandra استفاده می کند. با توجه به مقدار زیادی از داده ها در پروژه های فعلی ژنومیک ، به خصوص با عملکرد بالا نگران هستیم. ما نتایج خود را با یک سیستم رابطه ای PostgreSQL و یک سیستم پایگاه داده NoSQL دیگر ، MongoDB مورد بحث و مقایسه قرار می دهیم.

2- بسیاری از نوآوری های مرتبط در مدیریت داده ها ناشی شده است برنامه های وب 2.0. با این حال ، روش ها و ابزارها موجود در سیستم های رابطه ای ، ممکن است گاهی اوقات ، محدودیت آنها را ایجاد کند گسترش. بنابراین ، برخی از محققان تصمیم گرفته اند که راه حل های پایگاه داده خود را در مقیاس وب توسعه دهند. پایگاه های داده (NoSQL) نه تنها SQL به عنوان یک پایگاه داده شده اند راه حل برای مسائل مقیاس پذیری ذخیره سازی ، موازی سازی و مدیریت از حجم زیادی از داده های بدون ساختار.

به طور کلی ، سیستم های NoSQL ویژگی های زیر را دارند:

(i) آنها بر اساس یک مدل داده غیر مرتبط هستند.

(ii) آنها اعتماد می کنند.

(III) در پردازش توزیع شده در دسترس بودن و مقیاس پذیری بالانگرانی های اصلی است.

(v) برخی از آنها بدون برنامه هستند و دارای توانایی رسیدگی به داده های ساختار یافته و غیر ساختاری.

چهار پایگاه اصلی پایگاه داده NoSQL وجود دارد:

(i) ذخیره مقدار کلیدی: داده ها به عنوان مقادیر جفت کلید ذخیره می شوند. این سیستم ها مشابه دیکشنری ها هستند ، جایی که داده ها هستند توسط یک کلید واحد آدرس دهی می شود. مقادیر منزوی هستند و مستقل از دیگری ، و روابط با منطق برنامه اداره می شود.

(ii) پایگاه داده خانواده ستون: ساختار داده را تعریف می کند به عنوان یک ستون از پیش تعریف شده ستون های فوق العاده و ساختارهای ستونی را می توان در نظر گرفت.

(iii) ذخیره سازی مبتنی بر اسناد: فروشگاه اسناد از مفهوم ذخیره ارزش کلیدی. اسناد مجموعه‌ای هستند از ویژگی ها و ارزش ها ، جایی که یک ویژگی است می تواند چند ارزشی باشد. هر سند دارای یک شناسه است و کلید ، که در مجموعه منحصر به فرد است و شناسایی می کند.

(iv) پایگاه داده های نمودار: نمودارها برای نشان دادن استفاده می شوند و طرح ها یک پایگاه داده نمودار با سه انتزاع کار می کند: گره ، روابط بین گره ها و مقدار کلید جفتهایی که می توانند به گره ها و روابط متصل شوند.

– سیستم پایگاه داده کاساندرا. کاساندرا ابری است سیستم پایگاه داده ، مقیاس پذیر گسترده ، طراحی شده برای مقدار زیادی از داده ها را از چندین سرور ذخیره کنید ، در حالی که فراهم کردن داده های سازگار و در دسترس. مبتنی است در مورد معماری Amazon's Dynamo و همچنین در مدل داده BigTable Google. Cassandra را فعال می کند.

مانند یک مدل مقدار کلیدی ، که در آن هر ردیف یک ردیف منحصر به فرد دارد key ، یک ویژگی اقتباس شده از دینامو است. کاساندرا با استفاده از ویژگی ها یک پایگاه داده ترکیبی NoSQL در نظر گرفته می شود از هر دو پایگاه داده با ارزش کلیدی و ستون گرا. معماری کاساندرا از گره ها ، خوشه ها ، داده ها ساخته شده است. مراکز و یک تقسیم کننده گره یک نمونه فیزیکی از است کاساندرا. کاساندرا از معماری master-slave استفاده نمی کند. بلکه کاساندرا از معماری نظیر به نظیر استفاده می کند که همه گره ها برابر هستند.

خوشه گروهی از گره ها یا حتی a است تک گره گروهی از خوشه ها مرکز داده هستند. یک تقسیم کننده یک عملکرد هش برای محاسبه رمز هر کلید ردیف است. هنگامی که یک ردیف وارد می شود ، یک رمز بر اساس محاسبه می شود روی کلید ردیف منحصر به فرد آن. این رمز در کدام گره تعیین می کند آن ردیف خاص ذخیره خواهد شد. هر گره از یک خوشه است طیف وسیعی از داده ها را بر اساس یک توکن مسئول می کند. وقتی که ردیف درج می شود و توکن های آن محاسبه می شود ، این ردیف ذخیره می شود گره ای که مسئول این رمز است. مزیت اینجاست که چندین ردیف را می توان به طور موازی در پایگاه داده نوشت ، به عنوان مثال هر گره مسئول درخواست های نوشتن خود است. با این حال این ممکن است به عنوان یک اشکال در مورد استخراج داده ها دیده شود ، گلوگاه شدن پارتیشن بندی که از توکن ها برای اختصاص بخشهای مساوی از داده استفاده می کند به هر گره. این تکنیک انتخاب شده است زیرا هش سریع و عملکرد هش آن به توزیع یکنواخت کمک می کند داده ها به تمام گره های یک خوشه. عناصر اصلی کاساندرا فضای ستونی ، ستون هستند خانواده ها ، ستون ها و ردیف ها فضای کلیدی شامل مراحل پردازش تکثیر داده ها و شبیه به a است طرحواره در یک پایگاه داده رابطه ای. به طور معمول ، یک خوشه یکی دارد فضای کلید در هر برنامه. خانواده ستون مجموعه ای از مقدار کلید است جفتهایی که شامل یک ستون با کلیدهای ردیف منحصر به فرد آن هستند. a ستون کوچکترین افزایش داده است که شامل a است نام ، یک مقدار و یک مهر زمان. ردیف ها ستون هایی با همان کلید اصلی هستند هنگامی که یک عمل نوشتن رخ می دهد ، Cassandra بلافاصله دستورالعمل ها را در log ذخیره می کند ، که به دیسک سخت (HD)می رود. داده های حاصل از این عملیات نوشتن در حافظه ذخیره شده ذخیره می شود که در RAM می ماند. فقط با رسیدن به محدودیت حافظه از پیش تعیین شده ، این داده ها روی SSTable هایی که در HD می مانند ، نوشته می شوند. سپس ، ثبت

تعهدات و قابل یادداشت پاک می شود. در صورت عدم موفقیت در مورد جدول های یادبود ، کاساندرا دستورالعمل های کتبی موجود در گزارش کمیته را دوباره اجرا می کند. وقتی دستورالعمل استخراج اجرا می شود ، کاساندرا ابتدا اطلاعات موجود در جدول های یادآوری را جستجو می کند. یک RAM بزرگ مقدار زیادی از داده ها را در جدول های یادآوری و داده های کمتری را در HD ذخیره می کند ، در نتیجه دسترسی سریع به اطلاعات وجود دارد.

3. ذخیره داده های ژنومی ماندگاری داده های ژنومی یک مشکل اخیر نیست. در سال 2004، بلومند شارپ مشکلات مدیریت را شرح داد این داده ها. یکی از اصلی ترین مشکلات رشد بود تعداد داده تولید شده توسط سریال هاست. برای ذخیره داده های ژنومی در قالب FASTQ.بتمن و وود استفاده از NoSQL را پیشنهاد داده اند پایگاه های داده به عنوان یک جایگزین خوب برای ادامه داده های ژنتیکی. با این حال ، هیچ نتیجه عملی داده نمی شود. Ye and Li پیشنهاد استفاده از کاساندرا به عنوان یک سیستم ذخیره سازی. آنها چندین گره را در نظر بگیرید تا هیچ شکافی در سازگاری داده ها وانگ و تانگ برخی را نشان دادند دستورالعملهای ایجاد برنامه برای انجام داده ها عملیات در کاساندرا. Tudorica و Bucur برخی از NoSQL را مقایسه کردند پایگاه داده های پایگاه داده رابطه ایMySQL با استفاده از YCSB(معیار خدمات ابری یاهو). آنها نتیجه می گیرند که در محیطی که در آن عملیات نوشتن در MySQL غالب است تاخیر نسبت به کاساندرا مقایسه می شود. نتایج مشابهی در مورد بهبود عملکرد برای نوشتن در Cassandra ، وقتی با MS SQL مقایسه می شود Express ، همچنین توسط لی و مانوهاران گزارش شد. بسیاری از کارهای پژوهشی نتایج مربوط به آنها را ارائه می دهند عملکرد یک سیستم پایگاه داده Cassandra برای گسترده حجم داده ها در این مقاله ، ما تصمیم گرفته ایم که ارزیابی کنیم عملکرد سیستم پایگاه داده Cassandra NoSQL به طور خاص برای داده های ژنومی.