DHS: Adaptive Memory Layout Organization of Sketch Slots for Fast and Accurate Data Stream Processing

ACM 2021

2 citations

Tsinghua University

الگوریتم پردازش استریم DHS آیتم‌های ورودی را به باکت ها هش می‌کند و خانه[1](سلول‌ها یا همان شمارنده‌ها) درون هر باکت (اسلات حافظه) را به صورت پویا مدیریت می‌کند. در واقع سایز تمامی خانه‌ها درون یک باکت به صورت سازگاربا سایز و توزیع جریان واقعی تنظیم شده‌است. بدین صورت که جریان هر چه بزرگتر شد بیت‌های بیشتر و جریان‌های کوچک بیت‌های کمتر تخصیص می‌یابد. در ابتدا نیز تعداد بیت‌ها کم می‌باشد(اذا مبتنی بر اسکچ می‌باشد و از ویژگی نرخ گذر بالای آن استفاده می‌کند). همچنین یک مکانیزم کوئری longest fingerprint first برای تسریع بازیابی استفاده می‌کند.

دربحث ارزیابیْ، دقتِ بالا(روش‌های قبلی تخمینی می‌باشند)، نرخ گذر بالا(این که داده‌ها را با سرعتی حداقل برابر با سرعت تولید آنها پردازش کند)، عمومیت بالا و کارایی حافظه DHS را در مقایسه با دیگر راه‌حل‌های هیبریدی ارائه شده، برای مسائل مختلف دیتااستریمینگ(اندازه‌گیری) اثبات می‌کند(تمامی الگوریتم‌ها به زبان C پیاده‌سازی شده و در گیتهاب موجود می‌باشند):

- تخمین سایز جریان – k-top
- تخمین توزیع اندازه جریان: تعداد جریانهایی که یک سایز مشخص را دارند.
- تشخیص heavy hitter ها
- تشخیص آنومالی‌ها و تغییرات شاخص
- تشخیص آنتروپی: آنتروپی اندازه‌جریان

الگوریتم‌های محاسبات دیتا استریمینگ:

با توجه به خصوصیات دیتااستریم‌ها، داده را در حافظه‌های محدود به صورت موقت ذخیره می‌کنند، آنها را یک بار پردازش کرده و آمارش را در داده ساختارهایی برای یک کویری که آخر یک بازه یا در انتهای کار الگوریتم اجرا می‌شود، نگه‌داری می‌کنند. داده‌ساختارهای مورد استفاده در دیتااستریمینگ، بایستی یک تخمین کلی در مورد جریان‌ها ارائه دهند. هر آیتم ورودی استریم متعلق به یک جریان می‌باشد. در محاسبات دیتا استریمی با دو متد اصلی insert و query مواجه هستیم و برای مسائل استخراج داده[2] پیچیده‌تر، می‌توان توابعی افزون بر اینها و بر پایه اینها ایجاد کرد

هر کدام از الگوریتم‌های دیتااستریمینگ برای یک دسته از محاسبات مناسب(۵ دسته بالایی) می‌باشند و کارایی را در زمینه‌های کارایی حافظه، دقت یا نرخ گذر بهبود می‌بخشند. این روش‌ها را به سه دسته می‌توان تقسیم کرد:

- روش‌های مبتنی بر اسکچ: اطلاعاتی کلی از همه جریان‌ها را ذخیره می‌کنند و از نرخ گذربالایی برخوردار می‌باشند.
- روش‌های مبتنی بر شمارنده: اطلاعات را به صورت دقیق ذخیره کرده و دارای نرخ‌گذر پایین می‌باشند. آرایه‌ای از چندین اسلات که هر اسلات یک دوتایی (کلید،مقدار)(یا همان شمارنده) می‌باشد. کلید همان شناسه جریان و مقدار، فرکانس جریان می‌باشد. زمانی که یک جریان جدید می‌آید، جریان با کمترین فرکانس از آرایه خارج خواهد شد.

[1] cells
[2] Data minning

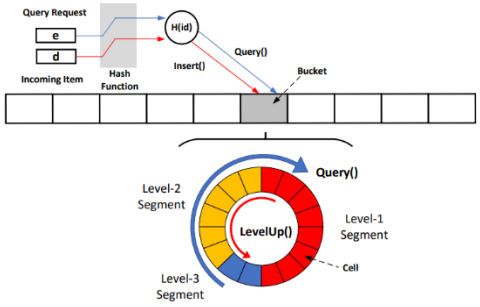
- روش‌های هیبریدی: ترکیبی از ایده‌های اصلی روش‌های بالا با یک مصالحه[3] بین دقت و نرخ‌گذر(هدف برخورداری از هردو مزیت می‌باشد) می‌باشند و برای مسائل بیشتر اندازه‌گیری مناسب می‌باشند. وبه جای مقدار فرکانس که در اسکچها استفاده می‌شد، از دوتایی (کلید،مقدار) استفاده می‌کنند.آرایه‌ای از باکت‌ها می‌باشند که هر باکت نیز شامل چندین جفت دودویی است. هر جریان به یک باکت هش می‌شود و (شناسه جریان، فرکانس جریان) در یک باکت ذخیره خواهد شد.

متود ما سه ایده کلی دارد:

- تخصیص سازگار حافظه برای شمارش فرکانس: در شمارنده‌ها و اسکچها برای جریان‌های ضعیف هم از مقدار حافظه معین تعریف شده، برای اسلات‌ها و ذخیره داده‌ها استفاده می‌کنند. اما درDHS ابتدا سایز کم می‌دهد و سپس با ورود جریان‌های جدید، می‌تواند افزایش یابد.
- تخصیص سازگار حافظه برای شناسه جریان: اسکچها شناسه جریان را نگه‌داری نمی‌کنند و منجر به خطای تخمین برای جریان‌های موشی می‌شوند. در طول پردازش آنلاین، زمانی که فرکانس یک جریان پایین است، اثر انگشت[4] آن با بیت های کمتری تخصیص داده می شود. همانطور که فرکانس آن افزایش می یابد، اثر انگشت آن در محدوده بزرگتری دوباره محاسبه می شود.
- تخصیص سازگار بخش‌های از حافظه به جریان‌های موشی و فیلی: در ابتدا همه جریان ها موشی اند اما با گذشت زمان برای جریان‌های فیلی نیاز به حافظه بیشتر برای ذخیره سازی شناسه و مقادیر فرکانس جریان داریم که این‌ها را می‌توانیم از موش‌ها بگیریم.

معماری و ساختار:

داده ساختار DHS: داده ساختار ما آرایه ای شامل B باکت که هر باکت شامل یک آرایه مدور[5] و متدیتا هست، می‌باشد. آرایه مدور به چندین سگمنت متوالی و سلسله مراتبی منطقی تقسیم شده‌است. هر سگمنت چندین خانه[6](سلول) پایه‌ای دارد که یک دوتایی (اثرانگشت، فرکانس) می‌باشد اما سایز این دوتایی بنا به سگمنت آن می‌تواند متغیر باشد. مرز بین سگمنت‌ها قابل تغییر می‌باشد و در متادیتای باکت ذخیره شده‌است. هر باکت یک آرایه ۳ سطحی مدور می‌باشد. فیلدهای اثرانگشت و فرکانس، تعداد بیت‌هایشان یکسان می‌باشد و برای سطوح ۱، ۲ و ۳ به ترتیب ۸، ۱۲و ۱۶ بیت می‌باشد. طول آرایه مدور(سایز باکت) عدد ثابت W می‌باشد. تابع هش h0 برای محاسبه اندیس جریان(0,B-1) در آرایه باکت می‌باشد.h1,h2,h3 اثر انگشت هر سطح را برمی‌گردانند و بازه های آنها به ترتیب (0-255)،(0-4095)و (0-65535) می‌باشد. در حین پردازش استریم، مرز هر سگمنت بنا به جریان داخلش تغییر می‌یابد، که در نهایت منجر به تنظیم خانه‌ها (سلول‌ها) در هر سطح می‌شود. باید توجه داشت که زمانی که یک باکت سرشار از خانه‌های سطح بالا باشد، هر باکت مثل روش شمارنده عمل می‌کند و اگر سرشار از خانه‌های سطح پایین باشد، همانند یک اسکچ عمل خواهد کرد.



طریقه کار: در ابتدا همه فیلدها برابر صفر و هر باکت تنها خانه‌های سطح پایین می‌باشد. شامل عملیات زیر می‌باشد:

- query: از روش longest fingerprint first استفاده می‌کند. ابتدا باکت متناظرش و سپس خانه‌ها را ازسطح پایین تا سطح بالا پیمایش می‌کند. خروجی توابع هش را بر اساس شناسه جریان محاسبه می‌کند.
- insert: پس از آن که باکت متناظر را پیداکرد، در آن باکت عمل query را اجرا می‌کند که حالتهای زیر ممکن است رخ دهد:

[3] Trade-off
[4] fingerprint
[5] Circular array
[6] cell

- اگر دریکی از سطح‌ها رکورد پیدا شد و فرکانس مورد نظر را افزایش میدهد با این که هیچ گونه سرریزی رخ نخواهد داد.
- اگر دریکی از سطح‌ها(k) رکورد پیدا شد اما در صورت درج، سرریز رخ می‌داد، از سطح 0 تا k برای یافتن سطحی که خانه آزاد کافی برای سطح جدید k+1 را داشته باشد، جستجو می‌کند. اگر جستجو موفق بود، خانه‌های آزادشده را به سطح k+1 منتقل می‌کند. اگر موفق نبود، با یک احتمالی[۷] از مقدار فرکانس کوچکترین خانه سطح k+1 می‌کاهد و اگر مقدار آن از فرکانس جریان ورودی کمتر بود، با جریان ورودی جایگزین خواهد شد.
- اگر query پاسخ منفی برگرداند، خانه جدید به سطح یک اضافه خواهد شد.
- اگر query پاسخ منفی برگرداند و سطح یک خانه جدید نداشت، با یک احتمالی[۷] از فرکانس کوچکترین خانه سطح ۱ می‌کاهد و اگر مقدار کوچکترین خانه به ۰ رسید، جریان ورودی را با آن جایگذاری می‌کند.

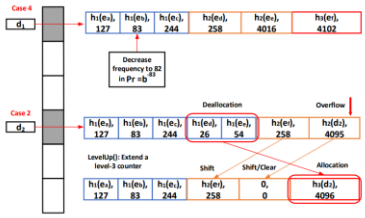


Figure 3: Insertion Examples of DHS

ارزیابی راهکار ارائه شده و مقایسه با دیگر راهکار:

پیش از مقایسه DHS با دیگر راهکارها، به بررسی میزان تاثیر پارامترهای تعداد باکت(B)، اندازه باکت(W) و پارامتر exponential decay (b) پرداخته و بهینه ترین آن‌ها را انتخاب می‌کند.

[7] Exponential decay parameter