دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات

پایان‌نامه کارشناسی
گرایش نرم‌افزار

عنوان

پیاده‌سازی یک ابزار داده‌کاوی مبتنی بر آپاچی اسپارک برای داده‌های جاری

نگارش
سینا شیخ‌الاسلامی

اساتید راهنما
دکتر سید رسول موسوی
دکتر امیرحسین پی‌براه

تیر ۱۳۹۵

اینجانب سینا شیخ‌الاسلامی متعهد می‌شوم که مطالب مندرج در این پایان نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است، مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

سینا شیخ‌الاسلامی

امضا

تقدیمات شاعرانه

تقدیر و تشکر:

تشکر رسمی

دکتر موسوی، دکتر پی‌براه، دکتر باقری

خانواده

فرزاد، ساسان، ریحانه

چکیده

کاوش و پردازش داده‌های جاری همواره بخش مهمی از پژوهش‌های مربوط به داده‌کاوی را به خود اختصاص داده است. با این وجود، با پیشرفت‌های اخیر در فناوری‌های رایانش ابری و سیستم‌های توزیع شده و همچنین فراگیر شدن استفاده از روش‌ها و ابزارهای تحلیل داده‌های حجیم، و به دلیل چالش‌ها و ویژگی‌های منحصر به فرد این دسته از داده‌ها، پژوهش در زمینه‌ی کاوش داده‌های جاری اهمیت روزافزونی یافته است.

در سالیان اخیر بسترهای مختلفی برای پردازش داده‌های حجیم و جریان‌داده‌ها تولید شده‌اند که از میان آن‌ها می‌توان به آپاچی اسپارک، آپاچی استورم، و آپاچی فلینک اشاره کرد. این بسترها دارای ابزارهایی برای پردازش داده‌های جاری هستند اما هنوز بسیاری از الگوریتم‌ها و روش‌های متداول کاوش داده‌های جاری برای استفاده برروی این بسترها پیاده‌سازی و آماده نشده‌اند.

هدف از این پروژه، پیاده‌سازی ابزاری متن‌باز برای کاوش داده‌های جاری می‌باشد. این ابزار شامل کتابخانه‌ای از الگوریتم‌های کاوش و پردازش داده‌های جاری، رابط کاربری گرافیکی برای مدیریت منابع جریان‌داده، تعریف و اجرای عملیات داده‌کاوی، نمایش نتایج حاصل از اجرا، مدیریت و نظارت بر محیط اجرای عملیات، و همچنین یک تولیدکننده‌ی جریان‌داده می‌باشد. این ابزار بر بستر آپاچی اسپارک و رابط برنامه‌نویسی اسپارک‌استریمینگ به عنوان یکی از بسترهای پیشرو برای پردازش داده‌های جاری و به طور کلی داده‌های حجیم پیاده‌سازی شده است.

در این پایان‌نامه، در ابتدا توضیحاتی در مورد چرایی نیاز به پردازش و کاوش داده‌های جاری، و چالش‌ها و مفاهیم کلیدی مرتبط با پردازش داده‌های جاری ارائه خواهد شد. سپس، راه‌کارهای موجود برای حل این مسأله و راه حل پیشنهادی مورد بحث قرار خواهد گرفت. در نهایت، به شرح پیاده‌سازی راه حل، نتایج حاصل شده، و کارهای آینده پرداخته می‌شود.

واژه‌های کلیدی:

داده‌های جاری، داده‌کاوی، جریان‌داده‌ها، داده‌های حجیم، الگوریتم‌های توزیع شده، آپاچی اسپارک

فهرست عنوان‌ها

صفحه

فهرست شکل‌ها

صفحه

فهرست جدول‌ها

صفحه

فصل اول -

مقدمه

پیشرفت‌های اخیر در حوزه‌ی سخت‌افزار منجر به این شده است که جمع‌آوری پیوسته‌ی داده‌ها به کاری آسان و متداول تبدیل شود. کارهای روزانه‌ای مانند جستجو در وب، ارسال پست در شبکه‌های اجتماعی، و خرید از فروشگاه‌های اینترنتی، به مرور حجم زیادی از داده تولید می‌کنند و با پردازش و کاوش این داده‌ها می‌توان به نتایج جالبی دست پیدا کرد. به عنوان مثالی دیگر، سامانه‌های کنترل خطوط حمل و نقل و ترافیک به طور معمول با جریان عظیمی از داده‌ها روبه‌رو هستند که تحلیل سریع آن‌ها می‌تواند در تصمیم‌گیری به مسئولین این حوزه‌ها کمک شایانی کند. همچنین، با تحلیل و کاوش کم‌تأخیر داده‌های مربوط به بسته‌های رد و بدل شده در یک شبکه‌ی کامپیوتری می‌توان به بروز ناهنجاری یا وقوع حملات خرابکارانه پی‌برد. در تمامی مثال‌های فوق، نوع خاصی از داده‌ها به نام داده‌های جاری مطرح هستند.

داده‌های جاری در مقایسه با دیگر انواع داده‌ها دارای خصوصیات منحصر به فردی هستند که پردازش و کاوش آن‌ها را به امری چالش‌برانگیز تبدیل می‌کند. از جمله‌ی این خصوصیات و چالش‌ها می‌توان به نیاز به الگوریتم‌های تک‌عبوره، نیاز به پردازش و کاوش کم‌تأخیر، عدم امکان ذخیره‌ی همه‌ی داده‌ها برروی حافظه‌های انبوه و پایگاه داده‌ها، امکان تغییر در نرخ ورود و حجم داده‌ها، و وقوع تحول در داده‌ها اشاره کرد.

در سالیان اخیر بسترهای مختلفی برای پردازش داده‌های حجیم ایجاد شده و توسعه یافته‌اند که از پردازش داده‌های جاری هم پشتیبانی می‌کنند. از جمله‌ی این بسترها می‌توان به آپاچی اسپارک، آپاچی استورم، و آپاچی فلینک اشاره کرد. با این حال و با وجود این که این بسترها دارای ابزارها و قابلیت‌هایی برای پردازش به‌صرفه‌ی جریان‌داده‌ها هستند، در حال حاضر تقریبا هیچ‌یک از الگوریتم‌های معمول کاوش داده‌های جاری برای استفاده در این بسترها پیاده‌سازی نشده‌اند. از طرف دیگر، استفاده و بهره‌گیری از امکانات و قابلیت‌های این بسترها نیازمند دانش و تجربه‌ی فراوان در حوزه‌های مختلفی از جمله رایانش ابری، سیستم‌های توزیع‌شده، الگوریتم‌های موازی، و داده‌کاوی می‌باشد.

هدف از این پروژه، طراحی و پیاده‌سازی ابزاری مبتنی بر بسترهای توزیع‌شده پردازش داده‌های حجیم برای کاوش داده‌های جاری است. این ابزار که SDMiner نام دارد، شامل:

- یک رابط کاربری گرافیکی برای تعریف کارهای[1] داده‌کاوی، مدیریت جریان‌داده‌های ورودی، و نمایش نتایج به کاربران کتابخانه‌ای از الگوریتم‌های معمول کاوش داده‌های جاری، و
- کتابخانه‌ای از الگوریتم‌های کاوش و پردازش داده‌های جاری، مانند نمونه‌برداری تصادفی

می‌باشد. از میان بسترهای مختلف پردازش داده‌های حجیم، بستر توزیع‌شده آپاچی اسپارک برای استفاده‌ی این ابزار انتخاب شده است.

در ادامه‌ی این پایان‌نامه و در فصل دوم، به چالش‌ها، روش‌ها و ابزارهای پردازش داده‌های جاری پرداخته خواهد شد. فصل سوم به موازی‌سازی و پیاده‌سازی الگوریتم نمونه‌برداری تصادفی بدون تبعیض به عنوان یکی از معمول‌ترین الگوریتم‌های کاوش داده‌های جاری می‌پردازد. در فصل چهارم طراحی و پیاده‌سازی ابزار SDMiner مورد بررسی قرار خواهد گرفت. بدین منظور، معماری کلی و توصیف اجزای مختلف سیستم، متدولوژی مهندسی نرم‌افزار به کار رفته در طراحی و پیاده‌سازی این پروژه، جزئیات پیاده‌سازی قسمت‌های مختلف ابزار، و نتایج حاصل از پیاده‌سازی بیان خواهد شد. در نهایت، فصل پنج به جمع‌بندی و کارهای آینده مرتبط با این پروژه خواهد پرداخت. همچنین، در قسمت پیوست، بخش‌هایی از پیاده‌سازی و راهنمایی برای کار و برنامه‌نویسی با استفاده از رابط برنامه‌نویسی اسپارک‌استریمینگ آورده شده است.

[1] Job

فصل دوم –

چالش‌ها، روش‌ها و ابزارهای پردازش و کاوش داده‌های جاری

در این فصل، مفاهیم پایه‌ی مطرح در پروژه، از جمله خصوصیات داده‌های جاری، چالش‌های پردازش و کاوش آن‌ها، و راه‌کارهای موجود مورد بررسی قرار خواهد گرفت.

۲.۱ داده‌های جاری و کاربردهای آن‌ها

پیشرفت‌های سخت‌افزاری در سالیان اخیر منجر به این شده است که جمع‌آوری پیوسته‌ی داده‌ها به کاری آسان و متداول تبدیل شود. کارهای روزانه‌ای مانند جستجو در وب، ارسال پست در شبکه‌های اجتماعی، و خرید از طریق فروشگاه‌های اینترنتی، به مرور حجم زیادی از داده تولید می‌کنند و با پردازش و کاوش این داده‌ها می‌توان به نتایج جالبی دست پیدا کرد. به عنوان مثالی دیگر، سامانه‌های کنترل خطوط حمل و نقل و ترافیک به طور معمول با جریان عظیمی از داده‌ها روبه‌رو هستند که تحلیل سریع آن‌ها می‌تواند به مسئولین در تصمیم‌گیری‌ها کمک کند. همچنین، با تحلیل و کاوش کم-تأخیر داده‌های مربوط به بسته‌های رد و بدل شده در یک شبکه‌ی کامپیوتری، می‌توان به بروز ناهنجاری یا وقوع حملات خرابکارانه پی برد.

در تمامی مثال‌های بالا، نوع خاصی از داده‌ها به نام «داده‌های جاری» مطرح هستند. یک تعریف فرمال؟

۲.۲ چالش‌های پردازش و کاوش داده‌های جاری

داده‌های جاری در مقایسه با اشکال دیگر داده دارای خصوصیات منحصر به فردی هستند که پردازش و کاوش آن‌ها را به امری چالش‌برانگیز تبدیل می‌کند. از جمله‌ی این خصوصیات و چالش‌ها می‌توان به موارد زیر اشاره کرد [۱][۲]:

۱ - نیاز به الگوریتم‌های تک-عبوره : با زیاد شدن حجم داده‌های جاری، پردازش بهینه‌ی داده‌ها به وسیله‌ی الگوریتم‌های چند-عبوره دیگر امکان‌پذیر نخواهد بود. بنابراین، الگوریتم‌ها و بسترها باید به گونه‌ای طراحی شوند که با یک بار عبور از داده‌ها، به نتایج مطلوب دست پیدا کنند.

۲ - نیاز به پردازش و کاوش کم-تأخیر: بسیاری از کاربردها نیازمند آن هستند که پردازش داده‌های جاری مرتبط با آن‌ها، به صورت بهنگام یا کم-تأخیر انجام شود. برای مثال، پست‌های مرتبط با یک خبر فوری در شبکه‌های اجتماعی فقط در بازه‌ی زمانی کوتاهی باارزش هستند. همچنین، در صورت وقوع

یک تصادف در یک بزرگراه، تصمیم‌گیری هرچه سریعتر به کاهش تبعات نامطلوب منجر خواهد شد. مواردی مانند نظارت پزشکی و تشخیص ناهنجاری و حملات در شبکه‌های کامپیوتری هم از این دست کاربردها هستند.

۳ - عدم امکان ذخیره‌ی همه‌ی داده‌ها بر روی حافظه‌های انبوه و پایگاه داده‌ها: با گذشت زمان، حجم داده‌ها ممکن است به قدری زیاد شود که عملا ذخیره‌سازی آن‌ها بر روی دیسک و حافظه‌های انبوه امکان‌پذیر نباشد. از طرف دیگر، به دلیل سربار زیاد دسترسی به حافظه‌های انبوه و دیسک‌ها، پردازش و کاوش کم-تأخیر داده‌ها نیازمند آن است که پردازش داده‌ها در حافظه‌ی اصلی صورت گیرد و نیاز به دسترسی به حافظه‌های انبوه به حداقل برسد. این نیازمندی همچنین سبب می‌شود که در بسیاری از موارد، سامانه‌های پردازشی به صورت توزیع‌یافته[۲] و مبتنی بر بسترهای رایانش ابری طراحی و پیاده‌سازی شوند.

۴ - امکان تغییر در نرخ ورود و حجم داده‌ها: سرعت و حجم ورود داده‌های یک جریان‌های داده هم ممکن است در طول زمان تغییر کند. برای مثال، نرخ ورود داده‌ها به یک سامانه‌ی کنترل ترافیک جاده‌های بین شهری، در روزهای عادی با تعطیلات بسیار متفاوت است.

۵ – وقوع تحول در داده‌ها : بسیاری از جریان‌های داده، در طول زمان دچار تحول می‌شوند. یک مثال خوب برای تحول، تغییر در توزیع کلاس‌های مختلف داده در جریان داده می‌باشد. بنابراین، الگوریتم‌های پردازش داده‌های جاری باید به گونه‌ای طراحی و پیاده‌سازی شوند که وقوع تحول در طول زمان باعث کاهش کارآیی آن‌ها نشود.

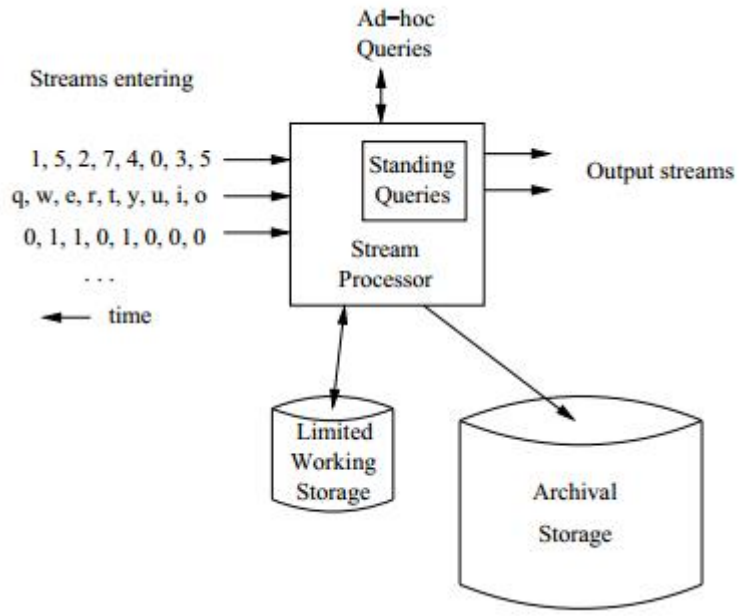
۲.۳ مدل کلاسیک پردازش داده‌های جاری

در شکل ۱ شمای کلی یک سامانه‌ی پردازش داده‌های جاری نشان داده شده است. مطابق شکل، جریانی از داده‌ها وارد سامانه شده و بخشی از آن‌ها که برای پردازش موردنیاز است در یک پایگاه داده محدود نگهداری می‌شوند. سپس پردازش‌های لازم روی این داده‌ها انجام‌شده و نتایج حاصل از آن در قالب یک

---

[۲] Distributed

جریان داده خروجی تولید می‌شوند. پردازشگر جریان بخش اصلی این سامانه است. در این پروژه، الگوریتم‌های پردازش و کاوش بر بستر یک پردازشگر جریان پیاده‌سازی خواهند شد.



شکل ۱- شمای کلی یک سامانه‌ی پردازش داده‌های جاری [۲]

با توجه به چالش‌های ذکر شده و نیازهای روزافزون به پردازش و کاوش داده‌های جاری، انتخاب بستری مناسب برای پیاده‌سازی الگوریتم‌های داده‌کاوی مورد نظر به تصمیمی مهم بدل می‌شود. در ادامه، مطرح‌ترین بسترهای موجود برای این کار معرفی و بررسی می‌شوند.

۲.۴ بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری

با توجه به خصوصیات منحصربه‌فرد داده‌های جاری – که در بخش ۲,۲ مورد بررسی قرار گرفت – پردازش داده‌های جاری در بسیاری از کاربردهای موردنظر نیازمند بسترهایی توزیع‌یافته می‌باشد. در همین راستا، در سالیان اخیر بسترهای توزیع‌یافته‌ی مختلفی تولید شده است که سه مورد از مهم‌ترین آن‌ها عبارتند از آپاچی فلینک ، آپاچی استورم ، و آپاچی اسپارک. در این قسمت، ابتدا معماری لایه‌ای معمول بسترهای توزیع‌یافته پردازش داده‌های جاری مورد بررسی قرار خواهد گرفت. در ادامه، با توجه به

معماری لایه‌ای معرفی شده مرور مختصری از هر یک از این سه بستر آورده می‌شود و پس از آن، بستر انتخابی برای پیاده‌سازی الگوریتم‌های داده‌کاوی در این پروژه – آپاچی اسپارک - و دلیل این انتخاب بیان خواهد شد.

۲.۴.۱ معماری عمومی بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری

شکل ۲ یک معماری لایه‌ای برای بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری را نشان می‌دهد. بسیاری از بسترهای مطرح پردازش داده‌های جاری (مانند آپاچی فلینک، آپاچی استورم، و آپاچی اسپارک) به نوعی برپایه‌ی این معماری توسعه یافته‌اند. مدل مورد بحث از چهار لایه‌ی رابط برنامه‌نویسی گراف کاربر، گراف اجرا، گره‌های اجرایی، و ارتباطات شبکه تشکیل شده است. همچنین یک پیکرپاره[3] برای مدیریت منابع گره‌های مختلف در این مدل به کار می‌رود. از جمله‌ی رایج‌ترین مدیرهای منابع می‌توان به Mesos و YARN اشاره کرد. در ادامه، لایه‌های مورد بحث به صورت اجمالی بررسی می‌شوند.

شکل ۲- معماری لایه‌ای بسترهای توزیع‌یافته پردازش داده‌های جاری

[3] Component

بالاترین لایه، رابط برنامه‌نویسی گراف کاربر است. این لایه، رابط برنامه‌نویسی کاربردی‌ای برای برنامه‌های کاربردی پردازش و کاوش جریان‌داده‌ها فراهم می‌کند. کاربران با استفاده از این رابط برنامه‌نویسی می‌توانند برنامه‌های کاربردی خود را به صورت گراف‌هایی مدل کنند که رأس‌های آن‌ها، گره‌های پردازشی هستند و رویدادها از طریق یال‌ها بین گره‌ها جریان پیدا می‌کنند.

لایه‌ی دوم، گراف اجرا نام دارد و در واقع گراف تبدیل‌یافته‌ای از گراف تعریف شده توسط کاربر (لایه‌ی اول) می‌باشد. تبدیل فوق توسط موتور پردازشی بستر و با توجه به محیط اجرایی صورت می‌گیرد و سپس گراف حاصل در خوشه‌ای از گره‌های پردازشی - لایه‌ی سوم - توزیع می‌شود.

لایه‌ی چهارم به مدیریت ارتباطات و شبکه‌ی بین گره‌های پردازشی مختلف – که ممکن است در خوشه‌های مختلفی قرار گرفته باشند – می‌پردازد. این لایه همچنین وظیفه‌ی سریالیزه کردن[۴] اشیاء و انتقال آن‌ها در شبکه با استفاده از پروتکل‌هایی مانند TCP، و کنترل جریانِ[۵] داده‌ها را برعهده دارد.

در نهایت، یک مدیر منابع وظیفه‌ی اداره‌ی منابع پردازشی مختلف، و زمان‌بندی وظایف[۶] میان خوشه‌ها و گره‌ها را برعهده دارد. بسیاری از بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری برای این منظور از برنامه‌های مدیریت منابعی مانند Mesos، YARN، و Nimbus استفاده می‌کنند.

موضوع مهم دیگری که در طراحی و استفاده از بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری مورد توجه قرار می‌گیرد، تضمین‌های پردازش[۷] و نحوه‌ی ترمیم پس از وقوع خرابی[۸] می‌باشد. در پردازش‌های توزیع‌یافته در مقیاس بزرگ، خطاها ممکن است به دلایل مختلفی، مانند خرابی گره‌ها، خرابی شبکه، اشکالات نرم‌افزاری، و محدودیت منابع رخ دهند[رفرنس به سروی جفری فاکس]. از آنجا که یکی از نیازمندی‌های پردازش داده‌های جاری، پردازش بهنگام یا کم‌تأخیر است، در صورت وقوع خرابی و خطا، سامانه پردازشی باید بتواند به سرعت خطا را رفع کرده و پردازش را ادامه دهد. همچنین، وقوع خطا

---

[۴] Serialization

[۵] Flow Control

[۶] Tasks

[۷] Processing Guarantees

[۸] Recovery from Failures

حتی‌الامکان نباید تأثیری در نتیجه‌ی پردازش داشته باشد. تضمین‌های پردازش، با توجه به نحوه‌ی ترمیم پس از وقوع خرابی در سامانه‌ی موردنظر تعریف می‌شوند.

به طور کلی، تضمین‌های پردازش در موتورهای پردازش جریان داده‌ها بر سه نوع هستند:

۱. دقیقا یک بار[۹]: این نوع از تضمین با ترمیم دقیق[۱۰] همراه است. پس از ترمیم دقیق، به جز افزایش مقطعی تأخیر، هیچ اثری از وقوع خرابی باقی نمی‌ماند و تمامی داده‌ها دقیقا یک بار پردازش می‌شوند.
۲. حداقل یک بار[۱۱]: این تضمین با ترمیم با عقبگرد[۱۲] متناظر است. در ترمیم با عقبگرد، هیچ بخشی از داده‌های جریان ورودی سامانه از بین نمی‌رود ولی وقوع خرابی ممکن است تأثیرات دیگری علاوه بر افزایش مقطعی تأخیر داشته باشد. در این صورت، ممکن است بعضی از داده‌ها دوباره (بیش از یک بار) پردازش شوند. به همین دلیل، این نوع تضمین، «حداقل یک بار» نام گرفته است.
۳. بدون تضمین[۱۳]: در صورت استفاده از روش ترمیم شکافی[۱۴] در صورت وقوع خرابی ممکن است بخشی از جریان ورودی به سامانه از بین برود. لذا در این حالت تضمینی برای پردازش همه‌ی داده‌ها وجود ندارد.

قسمت بعدی این فصل به معرفی و بررسی سه مورد از مطرح‌ترین بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری – آپاچی فلینک، آپاچی استورم، و آپاچی اسپارک - اختصاص دارد.

---

[۹] Exactly Once

[۱۰] Precise Recovery

[۱۱] At Leat Once

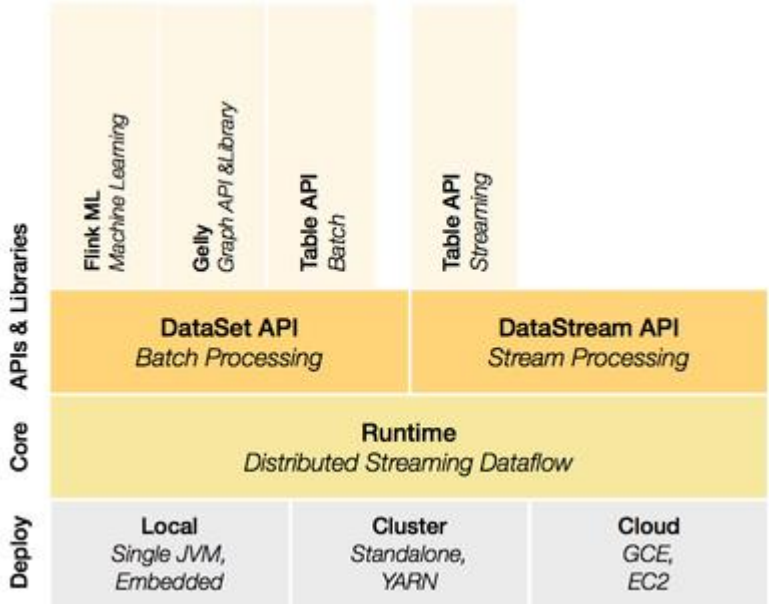
[۱۲] Rollback Recovery

[۱۳] No Guarantee

[۱۴] Gap Recovery

۲.۴.۲ آپاچی فلینک

آپاچی فلینک بستری توزیعی برای پردازش دسته‌ای داده‌های عظیم و داده‌های جاری است. این پروژه در سال ۲۰۱۰ در آلمان و با بودجه‌ی بنیاد تحقیقات آلمان و با نام استراتوسفر آغاز به کار کرد و از سال ۲۰۱۴ به عنوان یک پروژه‌ی سطح‌بالای بنیاد آپاچی مطرح شده است. فلینک برای پردازش داده‌های جاری، رابط برنامه‌نویسی نرم‌افزاری به نام DataStream API دارد که با استفاده از آن می‌توان به زبان‌های جاوا و اسکالا برنامه نوشت. شکل ۳، معماری آپاچی فلینک را نشان می‌دهد.



شکل ۳ - معماری لایه‌ای آپاچی فلینک

فلینک تضمین پردازش دقیقا یک بار را فراهم کرده و برای ترمیم پس از وقوع خرابی از حالت‌برداری[۱۵] استفاده می‌کند.

---

[۱۵] Checkpointing

۲.۴.۳ آپاچی استورم

آپاچی استورم یک بستر محاسباتی توزیع‌یافته و تحمل‌پذیرخطا برای محاسبات بهنگام در حجم وسیع است [۶] که بخش عمده‌ی آن به زبان برنامه‌نویسی کلوژر نوشته شده است. این پروژه ابتدا توسط تیمی در شرکت بک‌تایپ ایجاد شد. پس از مدتی شرکت توییتر این پروژه را خریداری کرده و آن را به صورت متن‌باز عرضه کرد. استورم از ماه سپتامبر سال ۲۰۱۴ به عنوان یک پروژه سطح بالای بنیاد آپاچی معرفی شده است.

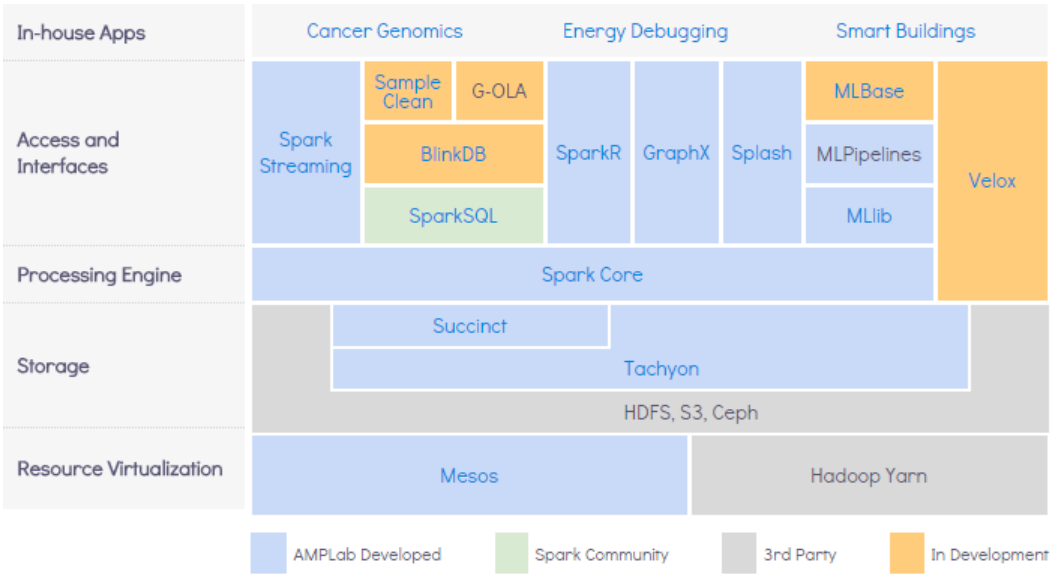
کاربران استورم می‌توانند به صورت صریح گراف‌های کاربری مورد نیاز خود را با تعریف گره‌ها، نحوه‌ی توزیع و ارتباط بین آن‌ها مشخص کنند. این مورد یکی از تفاوت‌های اصلی استورم با فلینک و اسپارک است، که در آنها امکان تعریف صریح گراف‌های کاربری وجود ندارد و خود موتور اجرایی با توجه به تعاریف سطح بالاتر صورت گرفته توسط کاربران این کار را انجام می‌دهد. استورم فاقد مکانیزم کنترل جریان است که این امر می‌تواند به ازدحام در میان‌گیرهای[۶] ورودی یا از دست رفتن داده‌های ورودی منجر شود. در زمینه‌ی تضمین‌های پردازشی، استورم تضمین حداقل یک بار پردازش را فراهم می‌کند.

۲.۴.۴ آپاچی اسپارک

آپاچی اسپارک یک موتور سریع وعام‌منظوره برای پردازش داده‌ها در مقیاس بزرگ است[۷][۸][۹]. اسپارک به عنوان موتور پردازشگر در استک تحلیل داده‌های برکلی مطرح می‌شود [۱۰] و می‌توان برای آن به زبان‌های جاوا، اسکالا، پایتون , و آر برنامه نوشت. اسپارک شامل تعدادی رابط برنامه‌نویسی نرم‌افزار برای پردازش داده‌های جاری (اسپارک‌استریمینگ) [۱۱]، کار با داده‌های ساختارمند (اسپارک سیکوئل)، یادگیری ماشین (ام‌ال‌لیب)، و پردازش گراف (گراف‌اکس) می‌باشد. در شکل ۴ محل قرارگیری اسپارک و رابط‌های برنامه‌نویسی آن در استک تحلیل داده‌های برکلی نشان داده شده است. اجزای آبی‌رنگ در ابتدا در آزمایشگاه امپ‌لب دانشگاه برکلی توسعه داده شده‌اند و اسپارک هم یکی از همین موارد است.

---

[۶] Buffers



شکل ۴ – استک تحلیل داده‌های برکلی

داده‌ساختار اصلی اسپارک برای کار با داده‌های حجیم و داده‌های جاری، مجموعه‌داده‌ی ارتجاعی توزیع‌یافته[۷](RDD) نام دارد. RDD ها مجموعه‌هایی تغییرناپذیر از اشیا و تحمل‌پذیرخطا هستند که بر روی یک خوشه توزیع شده‌اند. می‌توان گفت رابط برنامه‌نویسی کاربردی اسپارک بر مبنای تعریف RDD ها و استفاده از عملگرهای مخصوص آن‌ها توسعه داده شده است.

همانند فلینک و برخلاف استورم، کاربران اسپارک نمی‌توانند گراف کاربری را به صورت صریح تعریف کنند. در عوض، موتور اجرایی اسپارک با توجه به عملگرهای استفاده شده در برنامه‌ی کاربردی گراف موردنظر را ایجاد می‌کند.

در حقیقت، اسپارک یک موتور پردازش دسته‌ای[۸]داده‌ها است و در نتیجه برای پردازش داده‌های جاری، آن‌ها را به دسته‌های کوچکی تقسیم کرده و سپس اعمال پردازشی لازم را روی هر دسته انجام می‌دهد. به طور مشخص در مورد پردازش داده‌های جاری، یک جریان داده‌ی ورودی با داده‌ساختار دیگری به نام جریان گسسته‌شده[۹](DStream) متناظر می‌شود که در واقع دنباله‌ای از RDD هاست.

---

۷ Resilient Distributed Dataset

۸ Batch Processing

۹ Discretized Stream

RDDها از دو دسته اعمال پشتیبانی می‌کنند. دسته‌ی اول، تبدیل‌ها[۲۰]هستند که از یک RDD موجود، یک RDD جدید ایجاد می‌کنند. دسته‌ی دیگر اعمال، اقدام‌ها[۲۱]هستند که پس از انجام پردازش روی داده‌ها، یک مقدار را به عنوان خروجی به برنامه‌ی گرداننده[۲۲]بر می‌گردانند. برای مثال، یکی از معمول‌ترین تبدیل‌ها، تبدیل map می‌باشد که یک تابع را به تمامی اعضای یک مجموعه‌داده اعمال می‌کند و مجموعه‌داده‌ی جدیدی – که هر عضو آن، نتیجه‌ی اعمال تابع موردنظر بر اعضای مجموعه‌داده‌ی ورودی است - تولید می‌کند. count هم مثالی از یک اقدام است، که تعداد اعضای یک مجموعه‌داده را محاسبه کرده و به عنوان خروجی برمی‌گرداند.

لازم به ذکر است که تبدیل‌های موجود در اسپارک، اصطلاحا تنبل[۲۳]هستند، یعنی این تبدیل‌ها تا زمانی که برروی مجموعه‌داده نهایی اقدامی صورت نگیرد انجام نمی‌شوند. در عوض، زنجیره‌ی تبدیل‌هایی که برروی یک مجموعه‌داده اعمال می‌شوند در قالب دودمان[۲۴]مجموعه‌داده‌ی نهایی نگهداری می‌شود و زمانی که در برنامه قرار شود یک اقدام روی مجموعه‌داده‌ی نهایی صورت بگیرد، تبدیل‌های موردنظر واقعا انجام می‌شوند تا مجموعه‌داده‌ی نهایی در عمل ایجاد شده و اقدام موردنظر بتواند روی آن صورت گیرد. این کار به افزایش سرعت و بهینگی پردازش‌ها در اسپارک منجر می‌شود. همچنین، تحمل‌پذیری خطا و ترمیم پس از وقوع خرابی در اسپارک با استفاده از همین دودمان‌های نگهداری شده امکان‌پذیر می‌شود.

نوشتن برنامه‌های کاربردی و پیاده‌سازی الگوریتم‌های مختلف بر بستر آپاچی اسپارک نیازمند شناخت کامل مجموعه‌ی تبدیل‌ها و اقدام‌ها می‌باشد. در واقع، یک برنامه‌ی مبتنی بر اسپارک شامل زنجیره‌ای از تبدیل‌ها و اقدام‌های روی مجموعه‌داده‌هاست.

---

۲۰ Transformations

۲۱ Actions

۲۲ Driver Program

۲۳ Lazy

۲۴ Lineage

۲.۴.۵ انتخاب بستر مناسب برای پیاده‌سازی الگوریتم

در بخش‌های قبل، سه بستر آپاچی فلینک، آپاچی استورم، و آپاچی اسپارک مورد بررسی اجمالی قرار گرفته و از بعضی جنبه‌های فنی با یکدیگر مقایسه شدند. یکی از تصمیم‌های مهم برای تولید هر ابزاری، انتخاب بسترها و تکنولوژی‌های مورد استفاده برای پیاده‌سازی می‌باشد. در این بخش به دلیل انتخاب آپاچی اسپارک به عنوان بستر توزیع‌یافته پردازش داده‌های جاری و هسته‌ی اصلی ابزار SDMiner پرداخته می‌شود.

هرسه بستر ذکر شده به واسطه‌ی رابط‌های برنامه‌نویسی خود این امکان را به توسعه دهندگان می‌دهند که الگوریتم‌های مختلفی، از جمله الگوریتم‌های کاوش داده‌های جاری را بر روی آن‌ها پیاده‌سازی کنند. با توجه به اینکه هرسه بستر نسبتا جدید هستند، یک جنبه‌ی مهم در تصمیم‌گیری برای انتخاب بستر پیاده‌سازی، در دسترس بودن مستندات پیاده‌سازی و توسعه‌ی نرم‌افزار، و فعال‌بودن جامعه‌ی توسعه‌دهندگان می‌باشد. هرسه‌ی این بسترها به صورت متن‌باز و در گیت‌هاب موجود هستند. شکل ۵ تعداد تغییرات ایجاد شده در کد در هر هفته را دربازه‌ی هفته‌ی اول فوریه‌ی ۲۰۱۵ تا پایان ژانویه ۲۰۱۶ (زمان شروع پیاده‌سازی پروژه) برای هر سه بستر نشان می‌دهد.

شکل ۵ – تعداد تغییرات اعمال شده در کد در هر هفته برای هر بستر در بازه‌ی فوریه‌ی ۲۰۱۵ تا ژانویه ۲۰۱۶

همان‌طور که مشاهده می‌شود، تعداد تغییرات اعمال شده در این بازه برای پروژه‌ی اسپارک به مراتب از استورم و فلینک بیشتر است. جدول ۱ به مقایسه‌ی برخی ویژگی‌های دیگر این سه بستر که در روند توسعه‌ی نرم‌افزارهایشان نقش مهمی دارند می‌پردازد. این اطلاعات از صفحات پروژه‌ها در وب‌سایت بنیاد آپاچی و همچنین مخازن کد گیت‌هاب متناظرشان و آمارهای وب‌سایت stackoverflow.com استخراج شده‌اند.

جدول1 - مقایسه‌ی برخی ویژگی‌های مربوط به توسعه‌ی سه بستر (در تاریخ ۳۱ ژانویه ۲۰۱۶)

	آپاچی اسپارک	آپاچی استورم	آپاچی فلینک
توسعه‌دهندگان فعال	۷۹۷	۲۰۰	۱۵۴
سئوالات تگ‌شده در وب‌سایت stackoverflow.com	۹۹۰۰	۱۳۹۳	۲۰۸

با توجه به شکل ۵ و جدول ۱ و نظر به فعال‌تر بودن جامعه‌ی توسعه‌دهندگان و دردسترس‌تر بودن مستندات و منابع آموزشی، آپاچی اسپارک به عنوان بستر پیاده‌سازی الگوریتم در این پروژه انتخاب می‌شود.

در بخش بعدی، مرور مختصری از رابط برنامه‌نویسی کاربردی آپاچی اسپارک برای کار با داده‌های جاری (اسپارک‌استریمینگ) آورده شده است.

۲.۵ مروری بر رابط برنامه‌نویسی کاربردی اسپارک‌استریمینگ

اسپارک‌استریمینگ، رابط برنامه‌نویسی کاربردی آپاچی اسپارک برای پردازش داده‌های جاری است. شکل ۵ جریان کلی ورودی و خروجی داده‌ها در اسپارک‌استریمینگ را نشان می‌دهد. داده‌ها می‌توانند از منابع مختلفی مانند آپاچی کافکا، توییتر، HDFS، Flume، و سوکت‌های TCP وارد شوند و پس از پردازش، خروجی را می‌توان برروی فایل‌سیستم‌های مختلف (مانند HDFS) و پایگاه‌های داده ذخیره کرد یا برروی داشبوردهای مختلف نمایش داد. در این پروژه، جریان‌داده‌های ورودی با استفاده از سوکت‌های

TCP (به عنوان کلی‌ترین درگاه ورودی) خوانده می‌شوند و نتایج حاصل از وظایف داده‌کاوی بر روی یک داشبورد تحت وب نمایش داده می‌شود.

شکل ۶ – جریان کلی ورودی و خروجی در اسپارک‌استریمینگ

همانطور که در بخش ۲,۴,۴ بیان شد، اسپارک‌استریمینگ برای پردازش جریان‌داده‌ها، آن‌ها را به دسته‌های کوچکی تقسیم کرده و سپس اعمال پردازشی لازم را روی هر دسته انجام می‌دهد. شکل ۶ این موضوع را نشان می‌دهد.

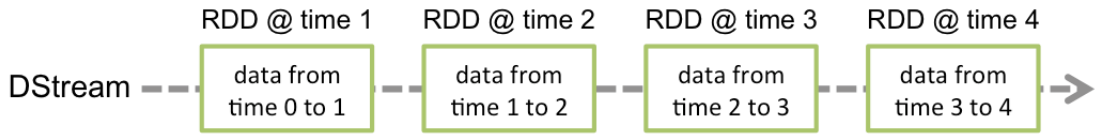
شکل ۷ – تقسیم جریان‌داده‌ی ورودی به دسته‌های داده در اسپارک‌استریمینگ

در اسپارک‌استریمینگ، برای پردازش داده‌های جاری، یک جریان داده‌ی ورودی با داده‌ساختار دیگری به نام جریان گسسته‌شده[۲۵](DStream) متناظر می‌شود که در واقع دنباله‌ای از RDD هاست. هر RDD

---

۲۵ Discretized Stream

موجود در یک DStream شامل داده‌های ورودی در یک بازه‌ی زمانی[۲۶]مشخص است. RDDهای موجود در DStream بر اساس بازه‌ی زمانی متناظرشان مرتب شده‌اند. شکل ۷ این مورد را بهتر نشان می‌دهد.



شکل ۸- جریان گسسته‌شده و RDDهای موجود در آن

رابط برنامه‌نویسی اسپارک‌استریمینگ (و به طور کلی اسپارک) امکان نوشتن برنامه به زبان‌های اسکالا، جاوا، آر، و پایتون را برای برنامه‌نویسان و توسعه‌دهندگان فراهم می‌کند. با توجه به اینکه اسکالا یک زبان برنامه‌نویسی تابعی[۲۷]می‌باشد، نحو آن به خوبی با چارچوب تفکر تابعی حاکم بر اسپارک (اِعمال زنجیره‌ای از تبدیل‌ها و اقدام‌ها روی مجموعه‌داده‌ها) هماهنگ است و به همین دلیل به عنوان زبان برنامه‌نویسی برای پیاده‌سازی الگوریتم‌های کاوش داده‌های جاری در این پروژه انتخاب شده است. لازم به ذکر است که اسکالا به طور کلی هم زبان اصلی مورد استفاده برای نوشتن برنامه‌های کاربردی مبتنی بر اسپارک بوده و پیاده‌سازی خود بستر آپاچی اسپارک نیز با استفاده از همین زبان صورت گرفته است.

۲.۶ خلاصه‌ی فصل

در این فصل مفاهیم پایه‌ی حوزه‌ی پردازش داده‌های جاری، چالش‌های این امر، و روش‌ها و معماری معمول بسترهای توزیع‌یافته برای پردازش و کاوش داده‌های جاری مورد بررسی قرار گرفت. سپس، سه بستر مطرح پردازش داده‌های جاری (آپاچی فلینک، آپاچی استورم، و آپاچی اسپارک) با توجه به مدل معرفی شده مورد بررسی قرار گرفتند و چگونگی انتخاب آپاچی اسپارک به عنوان بستر مورد استفاده در

---

[۲۶] Interval

[۲۷] Functional Programming Language

این پروژه شرح داده شد. در نهایت، به معرفی مختصری از رابط برنامه‌نویسی کاربردی آپاچی اسپارک برای کار با داده‌های جاری (اسپارک استریمینگ) پرداخته شد.

در فصل بعدی، الگوریتم نمونه‌برداری تصادفی بدون تبعیض، به عنوان یکی از معمول‌ترین الگوریتم‌های کاوش داده‌های جاری معرفی خواهد شد و چگونگی تبدیل آن به الگوریتمی توزیع‌یافته با توجه به چارچوب برنامه‌نویسی اسپارک‌استریمینگ مورد بررسی قرار خواهد گرفت.

فصل سوم –

الگوریتم نمونه برداری تصادفی توزیع‌یافته با مخزن ثابت

در این فصل به الگوریتم نمونه‌برداری تصادفی با مخزن ثابت[۲۸](RSFR) به عنوان یکی از معمول‌ترین الگوریتم‌های کاوش داده‌های جاری پرداخته خواهد شد. در ابتدا مبحث نمونه‌برداری[۲۹] و کاربردهای آن در وظایف کاوش داده‌های جاری مورد بررسی قرار خواهد گرفت. پس از آن، به چند روش معمول برای نمونه‌برداری اشاره شده و الگوریتم RSFR شرح داده خواهد شد. سپس، در مورد فرایند موازی‌سازی این الگوریتم و چگونگی طراحی و پیاده‌سازی نسخه‌ی توزیع‌یافته[۳۰](DSFR) بر بستر آپاچی اسپارک بحث خواهد شد.

۳.۱ نمونه‌برداری

نمونه‌برداری، یکی از روش‌های خلاصه‌سازی[۳۱] داده‌ها است. مسأله‌ی نمونه‌برداری عبارت است از انتخاب زیرمجموعه‌ای از داده‌ها به گونه‌ای که پاسخ‌های حاصل از پرس‌وجوهای[۳۲] صورت‌گرفته بر روی نمونه‌ی انتخاب شده به پاسخ‌های حاصل از پرس‌وجوهای صورت‌گرفته روی کل مجموعه داده‌ها نزدیک باشد. در صورتی که پرس‌وجوهای موردنیاز از قبل مشخص باشند می‌توان نمونه‌ها را با توجه به آن‌ها انتخاب کرد، اما در بسیاری از کاربردهای داده‌کاوی، پرس‌وجوهای تک‌کاره[۳۳] مطرح می‌شوند و به همین دلیل نمونه‌ی انتخاب شده باید دربرگیرنده‌ی تصویری کلی از مجموعه داده‌ها باشد. در مورد کاوش داده‌های جاری، با توجه به اینکه داده‌ها ممکن است در طول زمان دچار تحول شوند و حجم داده‌ها به نحوی است که نمی‌توان همه‌ی آن‌ها را در حافظه نگهداری کرد، انتخاب نمونه‌ی مناسب اهمیت بیشتری پیدا می‌کند.

---

۲ Random Sampling with a Fixed Reservoir

۲ Sampling

۳ Distributed Random Sampling with a Fixed Reservoir

۳ Synopsis Construction

۳ Query

۳ Ad-hoc

ویژگی اصلی نمونه‌برداری در مقایسه با سایر روش‌های خلاصه‌سازی داده‌ها – مانند تشکیل هیستوگرام[۳۴] یا موجک‌ها[۳۵] سادگی و به‌صرفه‌بودن آن است. با استفاده از روش‌های نمونه‌برداری می‌توان به سادگی به تصویری بدون تبعیض[۳۶] از کل مجموعه داده‌ها با تضمین خطای قابل اثبات[۳۷] دست پیدا کرد. همچنین، روش‌های دیگر خلاصه‌سازی برای داده‌های چندبعدی به راحتی قابل استفاده نیستند و درواقع در کاربردهایی که با داده‌های چندبعدی سر و کار دارند، پراستفاده‌ترین روش خلاصه‌سازی، نمونه‌برداری – و به طور مشخص نمونه‌برداری تصادفی – می‌باشد[کتاب آگاروال].

دو نمونه از پراستفاده‌ترین روش‌های نمونه‌برداری، نمونه‌برداری تصادفی با مخزن ثابت، و نمونه‌برداری مختصر[۳۸] می‌باشد. در این پروژه، الگوریتم نمونه‌برداری تصادفی با مخزن ثابت به عنوان اولین الگوریتم کتابخانه‌ی کاوش داده‌های جاری انتخاب، موازی‌سازی و پیاده‌سازی شده است. بخش بعدی به شرح این الگوریتم اختصاص دارد.

۳.۲ الگوریتم نمونه‌برداری تصادفی با مخزن ثابت (RSFR)

در این بخش، الگوریتم RSFR مورد بررسی قرار خواهد گرفت.

هدف از اجرای این الگوریتم، به دست‌آوردن نمونه‌ای بدون تبعیض با اندازه‌ی مشخص (مخزن ثابت) از جریان‌داده‌ی ورودی می‌باشد. شکل ۹ گام‌های اجرای این الگوریتم را نشان می‌دهد.

با فرض مخزن با سایز n، در ابتدا n عضو اول جریان‌داده در مخزن قرار داده می‌شوند. وقتی عضو k ام جریان‌داده وارد می‌شود، با احتمال n/k برای قرارگیری در مخزن انتخاب می‌شود. در صورت انتخاب شدن عضو k ام، چون سایز مخزن ثابت درنظر گرفته شده، برای قرارگیری در مخزن باید با یک عضو موجود در

---

۳۴ Histogram Construction

۳۵ Wavelets

۳۶ Unbiased

۳۷ Provable Error Guarantees

۳۸ Concise Sampling

مخزن جایگزین شود. عضوی که باید از مخزن خارج شود با احتمال 1/n از میان تمام اعضای فعلی مخزن انتخاب خواهد شد و سپس عضو k ام ورودی درجای آن قرار خواهد گرفت.

Random Sampling with a Fixed Reservoir (RSFR)

We have a reservoir of size **n**.
Input data comes in form of a stream of elements.

---

1 Add the first ***n*** elements of the data stream to the reservoir for initialization.

2 When the ***k***th element arrives, it is placed in the reservoir with a probability of ***n***/***k***.

3 If ***k***th element has to be added to the reservoir, an existing element of the reservoir with equal probability of 1/***n*** will be selected and removed from the stream, and the ***k***th element of input will replace it.

شکل ۹- گام‌های اجرای الگوریتم نمونه‌برداری تصادفی با مخزن ثابت

با استقرا برروی k می‌توان به این نتیجه رسید که خروجی الگوریتم RSFR یک نمونه‌ی بدون تبعیض از جریان‌داده است. اگر جریان‌داده برای مدت کافی ادامه پیدا کند، k خیلی بزرگ شده و به سمت بی‌نهایت میل می‌کند، پس تمامی اعضای جریان‌داده با احتمال یکسان می‌توانند در مخزن (نمونه) قرار بگیرند.

با وجود اینکه با استفاده از الگوریتم فوق می‌توان در هر زمان به نمونه‌ای بدون تبعیض از جریان‌داده‌ی ورودی دست پیدا کرد، در طراحی آن فقط یک گره درنظر گرفته شده که تمامی داده‌ها وارد آن می‌شوند، محاسبات در آن گره انجام می‌شود و نمونه را با توجه به محاسبات انجام شده تغییر می‌دهد. لذا نمی‌توان آن را برروی سیستم‌های توزیع‌یافته استفاده کرد و این با نیازمندی‌های کاوش و پردازش داده‌های جاری در تضاد است. از طرف دیگر، الگوریتم RSFR حساس به ورود عضو جدید جریان‌داده است، یعنی محاسبه احتمالات و به‌روزرسانی احتمالی مخزن با ورود هر عضو جدید به گره صورت می‌گیرد. این در حالی است که رابط برنامه‌نویسی اسپارک‌استریمینگ داده‌های ورودی را در بازه‌های زمانی ثابتی جمع‌آوری کرده و در قالب RDD قرار می‌دهد و انجام تبدیلات و اقدام‌ها فقط روی این RDDها امکان‌پذیر است و نمی‌توان در هر لحظه‌ای که یک عضو جدید وارد شد عملیات موردنظر را انجام داد.

در نتیجه، الگوریتم RSFR باید به گونه‌ای بازنویسی شود که هم در مدل برنامه‌نویسی اسپارک‌استریمینگ قابل پیاده‌سازی باشد و هم از رایانش توزیع‌یافته پشتیبانی کند. در این پروژه، الگوریتم RSFR برای برآورده کردن نیازمندی‌های فوق، بازطراحی شده و الگوریتم حاصل، «الگوریتم نمونه‌برداری تصادفی توزیع‌یافته با مخزن ثابت (DRSFR)» نام گرفته است.

در بخش بعدی الگوریتم DRSFR مورد بررسی قرار خواهد گرفت.

۳.۳ الگوریتم نمونه‌برداری تصادفی توزیع‌یافته با مخزن ثابت (DRSFR)

الگوریتم DRSFR در واقع توسعه‌ای از الگوریتم RSFR است که دو خصوصیت زیر را به آن اضافه می‌کند:

- موازی‌شده است و از رایانش توزیع‌یافته پیشتیبانی می‌کند، و
- با توجه به مدل برنامه‌نویسی اسپارک‌استریمینگ که داده‌های ورودی را به صورت دسته‌ای در بازه‌های زمانی مشخص در اختیار برنامه‌ی کاربردی (گرداننده) قرار می‌دهد قابل پیاده‌سازی است.

در این بازطراحی منطق عملکرد الگوریتم RSFR تغییری نمی‌کند و در نتیجه اثبات بدون تبعیض بودن این الگوریتم که در بخش ۳,۲ ذکر شد به قوت خود باقی است.

همان‌طور که در بخش قبلی توضیح داده شد، الگوریتم RSFR، احتمالات مورد نظر را محاسبه کرده و در صورت لزوم اقدام به به‌روزرسانی مخزن (نمونه) می‌کند. اسپارک‌استریمینگ داده‌های ورودی را در قالب RDD هایی که مربوط به بازه‌های زمانی مشخص هستند در اختیار برنامه‌ی کاربردی قرار می‌دهد (شکل ۸). ترتیب قرارگیری داده‌ها در RDDها همان ترتیب ورودشان به سامانه است ولی داده‌ها بر این اساس اندیس‌گذاری[۳] نمی‌شوند.

بسیاری از تولیدکننده‌های جریان‌داده (برای مثال سنسورهای سنجش کیفیت هوا) خود اقدام به زدن برچسب زمانی[۴] یا شماره‌گذاری داده‌های ورودی (بر اساس زمان و ترتیب تولید داده‌ها) می‌کنند، ولی خروجی دسته‌ی دیگر تولیدکننده‌های جریان‌داده فقط داده‌ی تولید شده (بدون برچسب و شماره) است.

---

[۳] Indexing

[۴] Timestamping

یک راه‌حل جامع باید بتواند از هر دو دسته‌ی داده‌های ورودی پشتیبانی کند، پس باید بتوان داده‌های ورودی را به ترتیب ورودشان شماره‌گذاری کرد تا محاسبه‌ی احتمالات مربوط به الگوریتم RSFR امکان‌پذیر شود، چون موتور پردازشی اسپارک بسته به تعداد گره‌های اجرایی، داده‌های ورودی را بین آن‌ها پخش می‌کند و ممکن است در بازگشت به برنامه‌ی گرداننده، ترتیب داده‌ها به هم بخورد.

در این پروژه، دو پیاده‌سازی برای الگوریتم DRSFR ارائه شده است. پیاده‌سازی اول مربوط به حالتی است که داده‌های ورودی از مبدا به ترتیب تولید شدنشان (و در نتیجه ترتیب ورودشان) شماره‌گذاری شده‌اند و به شماره‌گذاری مجدد نیازی نیست. پیاده‌سازی دوم، حالت کلی را که در آن داده‌ها بدون شماره‌گذاری هستند را پوشش می‌دهد و با استفاده از روش MapWithState اسپارک‌استریمینگ، داده‌ها به ترتیب ورودشان به سیستم اندیس‌گذاری می‌شوند. در ادامه به جزئیات هر دو پیاده‌سازی پرداخته می‌شود.

۳.۳.۱ پیاده‌سازی DRSFR برای داده‌های جاری شماره‌گذاری شده

پیش‌فرض این پیاده‌سازی این است که داده‌های ورودی به صورت رشته‌ای متشکل از مقدار اصلی و اندیس هستند. شکل ۱۰ گام‌های اجرای این الگوریتم را نمایش می‌دهد.

در انتهای هر بازه‌ی زمانی، DStream ای حاوی داده‌های ورودی در آن بازه تولید می‌گردد. کار با تبدیل رشته‌ی ورودی متناظر با هر عضو DStream به یک زوج مرتب (value, index) آغاز می‌شود. به عبارت دیگر، index در نقش k مورد بحث در الگوریتم RSFR می‌باشد و حال می‌توان از آن برای محاسبه‌ی احتمالات موردنیاز استفاده کرد.

در مرحله‌ی بعد، یک عمل صافی[۴۱] بر روی زوج‌مرتب‌ها (که به ترتیب ورودشان در RDD قرار گرفته‌اند) انجام می‌شود تا فقط زوج‌مرتب‌هایی باقی بمانند که شرط احتمالاتی موردنظر را برآورده می‌کنند. (لازم به ذکر است که برای n عضو اول، شرط احتمالی موردنظر همواره برقرار است، پس نیازی به قدم جداگانه‌ای برای درج اعضای اولیه نمونه نیست). با توجه به اینکه تعداد زوج‌مرتب‌های باقی‌مانده بسیار کمتر از کل مجموعه داده‌هاست، می‌توان عملیات درج آن‌ها در مخزن (نمونه) را برروی گره برنامه‌ی گرداننده انجام

---

[۴۱] Filter

داد. بدین منظور، از یک عمل جمع‌آوری[۲]استفاده می‌شود تا دسته‌های داده‌ی باقی‌مانده‌ی پخش شده

Distributed Random Sampling with a Fixed Reservoir (DRSFR)

For Pre-Indexed Data Streams and Apache SparkStreaming

We have a reservoir of size **n**.

random is a floating point number between 0 and 1, randomly generated on each occurrence.
Input data comes in form of a stream of pre-indexed elements. Indices are based on the order of production of data elements.
Input elements are in the form of: **"value, index"**

---

foreach batch interval ***interval***, do the following:

1 **transform** the ***interval.DStream*** into another DStream by applying a **map** operation to the containing RDD, resulting in a new DStream containing a RDD with its elements in the form of (value, index), and name the new DStream as ***indexedDStream***.

2 **filter** ***indexedDStream*** **foreach** element ***e*** in ***indexedDStream*** that satisfies the following predicate, and name the filtered DStream as ***filteredDStream***:

(***n*** / ***e***.index) > ***random***

3 **foreach** RDD ***r*** in ***filteredDStream***, do the following:

3.1 **collect** ***r*** inside an array named ***updateSet***.

3.2 **sort** ***updateSet*** in ascending order based on indices of elements of ***r***, call the new array as ***sortedUpdateSet.***

3.3 **foreach** element ***el*** in ***sortedUpdateSet***, do the following:

3.3.1 ***replaceIndex*** = a random integer between 0 and ***n***-1.

3.3.2 replace the reservoir element with the index of ***replaceIndex*** with ***el.***

شکل ۱۰- گام‌های اجرای الگوریتم DRSFR برای داده‌های شماره‌گذاری شده

در گره‌های مختلف، در یک مجموعه در گره اصلی (برنامه‌ی گرداننده) گردآوری شوند. سپس عملیات مرتب‌سازی صعودی این مجموعه بر اساس اندیس اعضا صورت می‌گیرد، و در نهایت با شروع از اول

مجموعه‌ی مرتب شده، به ازای هر عضو این مجموعه، یک عضو قدیمی موجود در مخزن از آن خارج شده و عضو جدید موردنظر در جای آن قرار می‌گیرد.

۳.۳.۲ پیاده‌سازی DRSFR برای داده‌های جاری بدون شماره

تفاوت این حالت با حالت قبل در این است که داده‌های ورودی به صورت رشته‌هایی فقط حاوی مقدار اصلی (و بدون شماره) هستند، و با توجه به پخش شدن داده‌ها بر روی گره‌های مختلف (پس از ورود)، باید نسبت به شماره‌گذاری ترتیبی آن‌ها اقدام نمود. گام‌های اجرای این نسخه از الگوریتم با حالت قبلی فقط در گام شماره‌ی یک (شکل ۱۰) تفاوت دارد. در این حالت، با استفاده از مفهوم وضعیت[۴۳]در اسپارک‌استریمینگ، هر عضو با شماره‌ی ورودش متناظر شده و اندیس‌گذاری می‌شود. باقی مراحل الگوریتم مانند شکل ۱۰ خواهد بود. این نسخه از الگوریتم، کلی‌ترین حالت موجود است و برای تمامی داده‌هایی که می‌توانند به صورت رشته‌ای وارد سامانه شوند (کاراکترها، مقادیر عددی و ...) قابل استفاده است.

۳.۴ خلاصه‌ی فصل

در این فصل ابتدا به کاربردهای نمونه‌برداری در وظایف کاوش و پردازش داده‌های جاری پرداخته شد. سپس به تعدادی از الگوریتم‌ها و روش‌های نمونه‌برداری و خلاصه‌سازی جریان‌داده‌ها اشاره شد و الگوریتم نمونه‌برداری تصادفی با مخزن ثابت (RSFR) به عنوان یکی از معمول‌ترین الگوریتم‌های نمونه‌برداری جریان‌داده‌ها مورد بررسی قرار گرفت. در نهایت نیز به طراحی و پیاده‌سازی نسخه‌ی توزیع‌یافته‌ی الگوریتم (DRSFR) با توجه به مدل برنامه‌نویسی اسپارک‌استریمینگ پرداخته شد. الگوریتم DRSFR اولین الگوریتم پیاده‌سازی شده در کتابخانه‌ی الگوریتم‌های داده‌کاوی این پروژه است.

---

[۴] Collect

[۴] State

فصل بعدی به طراحی و پیاده‌سازی ابزار مبتنی بر آپاچی اسپارک برای کاوش داده‌های جاری (SDMiner)، متدولوژی مهندسی نرم‌افزار به کار رفته در این پروژه، و نتایج حاصل از پیاده‌سازی اختصاص خواهد داشت.

فصل چهارم –

طراحی، پیاده‌سازی و ارزیابی سامانه

این فصل به

در ابتدا معماری کلی سامانه بیان شده و سپس به بحث در مورد هر یک از اجزای این ابزار پرداخته می‌شود. در خلال بررسی اجزای مختلف، به فناوری‌های مورد استفاده در پیاده‌سازی آن‌ها اشاره می‌شپس از آن،

فصل پنجم -
جمع‌بندی و کارهای آینده

منابع و مراجع

پیوست

بخش‌هایی از پیاده‌سازی

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Engineering and Information Technology

B.Sc. Thesis in Software Engineering

Title

SDMiner: A Tool for Mining Data Streams on Top of Apache Spark

By
Sina Sheikholeslami

Advisors
Dr. Seyyed Rasool Moosavi
Dr. Amir H. Payberah

June 2016