

دانشگاه صنعتي امیرکبیر  
(پلی تکنیک تهران)  
دانشكده 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

اينجانب سینا شیخ‌الاسلامی متعهد مي‌شوم كه مطالب مندرج در اين پايان نامه حاصل كار پژوهشي اينجانب تحت نظارت و راهنمايي اساتيد دانشگاه صنعتي اميركبير بوده و به دستاوردهاي ديگران كه در اين پژوهش از آنها استفاده شده است، مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذكر گرديده است. اين پایان نامه قبلاً براي احراز هيچ مدرك هم‌سطح يا بالاتر ارائه نگرديده است.

در صورت اثبات تخلف در هر زمان، مدرك تحصيلي صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پيگيري قانوني خواهد داشت.

كليه نتايج و حقوق حاصل از اين پایان نامه متعلق به دانشگاه صنعتي اميركبير مي‌باشد. هرگونه استفاده از نتايج علمي و عملي، واگذاري اطلاعات به ديگران يا چاپ و تكثير، نسخه‌برداري، ترجمه و اقتباس از اين پایان نامه بدون موافقت كتبي دانشگاه صنعتي اميركبير ممنوع است.   
نقل مطالب با ذكر مآخذ بلامانع است.

سینا شیخ‌الاسلامی

امضا

تقدیمات شاعرانه

**تقدیر و تشکر:**

آدمیزادی و رسمی

امیرحسین - خانواده

فرزاد، ساسان، ریحانه

چكيده

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 ادامه، توضیحاتی در مورد چرایی نیاز به پردازش و کاوش داده‌های جاری، و چالش‌ها و مفاهیم کلیدی مرتبط با آن ارائه خواهد شد. سپس، راه‌کارهای موجود برای حل این مسأله و راه حل پیشنهادی مورد بحث قرار خواهد گرفت. در نهایت، شرح پروژه‌ و مدل فرآیند طراحی نرم‌افزار مورد بررسی قرار می‌گیرد.

واژه‌های کلیدی:

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

|  |  |
| --- | --- |
| فهرست عنوان‌ها | صفحه |

[1 فصل اول - مقدمه 1](#_Toc454739378)

[2 فصل دوم - چالش‌ها، روش‌ها و ابزارهای پردازش و کاوش داده‌های جاری 4](#_Toc454739379)

[2‌.1‌ داده‌های جاری و کاربردهای آن‌ها 5](#_Toc454739380)

[2‌.2‌ چالش‌های پردازش و کاوش داده‌های جاری 5](#_Toc454739381)

[2‌.3‌ مدل کلاسیک پردازش داده‌های جاری 6](#_Toc454739382)

[2‌.4‌ بسترهای پردازش داده‌های جاری 7](#_Toc454739383)

[2‌.5‌ معماری عمومی بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری 8](#_Toc454739384)

[2‌.6‌ آپاچی فلینک 9](#_Toc454739385)

[3 فصل سوم - موازی‌سازی الگوریتم نمونه برداری تصادفی بدون تبعیض 10](#_Toc454739386)

[4 فصل چهارم - طراحی، پیاده‌سازی و ارزیابی سیستم 12](#_Toc454739387)

[5 فصل پنجم - جمع‌بندی و کارهای آینده 13](#_Toc454739388)

[منابع و مراجع 14](#_Toc454739389)

[پیوست 15](#_Toc454739390)

|  |  |
| --- | --- |
| فهرست شکل‌ها | صفحه |

[شکل 1- شمای کلی یک سامانه‌ی پردازش داده‌های جاری ]۲[ 7](#_Toc454739363)

[شکل 2- معماری لایه‌ای بسترهای توزیع‌یافته پردازش داده‌های جاری 8](#_Toc454739364)

|  |  |
| --- | --- |
| فهرست جدول‌ها | صفحه |

**No table of figures entries found.**

# 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]](#footnote-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.

# 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 برای کار با داده‌های جاری مورد بررسی قرار خواهد گرفت.

## 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؟

## 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 صورتی بدون تأخیر 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 رخداد 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[[2]](#footnote-2) 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.

## 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 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 خواهد شد.



شکل 1- 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.

## بسترهای پردازش داده‌های جاری

در سالیان اخیر برای پردازش داده‌های جاری، بسترهای 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 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.

## معماری عمومی بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری

شکل ۲ یک معماری لایه‌ای برای بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری را نشان می‌دهد. اکثر بسترهای مطرح پردازش داده‌های جاری به نحوی برپایه‌ی این معماری توسعه‌ یافته‌اند. (این ادعا را چگونه ثابت کنم؟‌رفرنس بدم به مقاله‌هه؟ یا دقیقن همین سه‌تا رو مثال بزنم، یعنی بگم که همانطور که در ادامه خواهیم دید مثلن؟ :دی). مدل مورد بحث از چهار لایه‌ی رابط برنامه‌نویسی گراف کاربر، گراف اجرا، گره‌های اجرایی، و ارتباطات شبکه تشکیل شده است. همچنین یک پیکرپاره[[3]](#footnote-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 2- معماری لایه‌ای بسترهای توزیع‌یافته پردازش داده‌های جاری

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 است. تبدیل فوق توسط موتور پردازشی 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[[4]](#footnote-4) اشیاء و انتقال آن‌ها در شبکه با استفاده‌ از پروتکل‌هایی مانند TCP را برعهده دارد.

در نهایت، یک مدیر منابع وظیفه‌ی اداره‌ی منابع پردازشی مختلف، و زمان‌بندی وظایف[[5]](#footnote-5) میان خوشه‌ها و گره‌ها را برعهده دارد. بسیاری از بسترهای توزیع‌یافته‌ی پردازش داده‌های جاری برای این منظور از نرم‌افزارهایی مانند Mesos، YARN، و Nimbus استفاده می‌کنند.

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.

## آپاچی فلینک

# فصل سوم - موازی‌سازی الگوریتم نمونه برداری تصادفی بدون تبعیض

# فصل چهارم - طراحی، پیاده‌سازی و ارزیابی سیستم

# 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

1. Job [↑](#footnote-ref-1)
2. Distributed [↑](#footnote-ref-2)
3. Component [↑](#footnote-ref-3)
4. Serialization [↑](#footnote-ref-4)
5. Tasks [↑](#footnote-ref-5)