

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](#_Toc454732633)

[2 فصل دوم - چالش‌ها، روش‌ها و ابزارهای پردازش و کاوش داده‌های جاری 4](#_Toc454732634)

[3 فصل سوم - موازی‌سازی الگوریتم نمونه برداری تصادفی بدون تبعیض 5](#_Toc454732635)

[4 فصل چهارم - طراحی، پیاده‌سازی و ارزیابی سیستم 7](#_Toc454732636)

[5 فصل پنجم - جمع‌بندی و کارهای آینده 8](#_Toc454732637)

[منابع و مراجع 9](#_Toc454732638)

[پیوست 10](#_Toc454732639)

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

**No table of figures entries found.**

|  |  |
| --- | --- |
| 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

# 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)