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 امیرحسین پی‌براه

دکتر سید رسول موسوی

خرداد 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:

آدمیزادی و رسمی

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 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

صفحه

فهرست عنوان‌ها

۱ فصل اول مقدمه..... ۱
۲ فصل دوم چالش‌ها، روش‌ها و ابزارهای پردازش و کاوش داده‌های جاری..... ۳
۳ فصل سوم طراحی، پیاده‌سازی و ارزیابی سیستم..... ۴
۴ فصل چهارم جمع‌بندی و کارهای آینده..... ۵
منابع و مراجع..... ۶
پیوست..... ۷

فهرست شکل‌ها **صفحه**

No table of figures entries found.

صفحه

فهرست جدول‌ها

No table of figures entries found.

فصل اول

مقدمه

در اینجا مقدمه را خواهیم نوشت

فصل دوم

چالش‌ها، روش‌ها و ابزارهای پردازش و کاوش داده‌های جاری

فصل سوم

طراحی، پیاده‌سازی و ارزیابی سیستم

فصل چهارم
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. Amir H. Payberah
Dr. Seyyed Rasool Moosavi

May 2016