طراحی و آزمایش داشبورد بلادرنگ تحلیل متن شبکهٔ اجتماعی توییتر

سعید روحانی*، طاهره پزشکی و بابک سهرابی
گروه مدیریت فناوری اطلاعات، دانشگاه تهران، تهران، ایران

چکیده

یکی از مباحث پژوهشی مهم امروز در حوزه فناوری اطلاعات و فناوری استفاده از دانش نهفته در داده‌هایی است که امروزه با سرعت بالا، حجم زیاد و با تنوع فراوان در فرمت داده تولید می‌شوند. داده‌هایی با چنین ویژگی‌هایی را کلان‌داده می‌نامند. استخراج، پردازش و بصری‌سازی نتایج حاصل از کلان‌داده امروزه به یکی از دغدغه‌های دانشمندان علم داده تبدیل شده است. گفتنی است که امروزه زیر ساخت‌ها، روش‌ها و ابزارهای بسیاری برای تحلیل کلان‌داده توسعه یافته‌اند. هدف این مقاله ارائه راهکاری برای استخراج و بصری‌سازی داده‌های شبکه اجتماعی توییتر به‌صورت بلادرنگ با حذف پایگاه‌های داده به‌عنوان نمونه‌ای از تحلیل کلان‌داده است. در این پژوهش یکی از راه حل‌های بصری سازی بلادرنگ، با استفاده از داده‌های توییتر به‌عنوان جریان ورودی، از آپاچی استورم به‌عنوان پلتفرم پردازشی و از D3.js برای نمایش داده‌ها ارائه خواهد شد؛ درنهایت داشبورد طراحی شده با استفاده از روش طراحی آزمایش‌ها و آزمون‌های آماری از نظر زمان طی‌شده برای پاسخ (Latency)، در انواع پیکره‌بندی‌های مختلف آپاچی استورم مورد ارزیابی قرار گرفته و درنهایت بلادرنگ‌بودن با میانگین زمان پاسخ برابر یک دقیقه و سی ثانیه تأیید شد.

واژگان کلیدی: کلان‌داده، بصری‌سازی، داشبورد بلادرنگ

Design and Test of the Real-time Text mining dashboard for Twitter

Saeed Rouhani*, Tahereh Pezeshki & Babak Sohrabi
Department of Information Technology Management, University of Tehran, Tehran, Iran

Abstract

One of today's major research trends in the field of information systems is the discovery of implicit knowledge hidden in dataset that is currently being produced at high speed, large volumes and with a wide variety of formats. Data with such features is called big data. Extracting, processing, and visualizing the huge amount of data, today has become one of the concerns of data science scholars.

The impact of big data on information analysis can be traced to four different parts. The first part is data extraction and processing, the second part is data analysis, the third part is data storage, and finally the visualization of the data. In the field of big data processing, in various studies, different categories have been presented. For example, in the studies of Hashim et al., big data processing is divided into two categories. These two types are: batch and real time. These two categories of processing, which nowadays are standard in any comprehensive big data solution, also have been introduced in Abawajy studies: batch processing is related to offline processing, and real-time processing is usually used to analyze the streaming data without any need to storage of data on disk. As data flows from various sources, the data is analyzed and processed real time, for immediate insight. As today's world is rapidly changing and survival in today's competitive

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال: ۹۶/۱۰/۱۴ ● تاریخ پذیرش: ۹۸/۰۷/۱۳ ● تاریخ انتشار: ۹۸/۱۲ ۲۸

[DOI: 10.29252/jsdp.16.4.151]

world requires instant decision-making based on flows of data, streaming data analysis is becoming increasingly important.

On the other hand, one of the great valuable sources of streaming data is the data generated by social networks' users such as Twitter. Social networks data sources are very rich sources for analysis as they come from the opinions and opinions of their users.

As discussed earlier, and since previous studies such as Flash's studies have focused more on batch analysis (offline data), this study has attempted to investigate a variety of tools and infrastructures related to big streaming data, and finally design a real-time dashboard based on Twitter social network streaming data.

The following article addresses two research questions: 1) How to design and implement a real-time dashboard based on social networks data? 2) Which different configurations are best suited for real-time dashboard analysis and visualization?

In other words, the purpose of this article is to provide a solution for extracting and visualizing Twitter's social network streaming data by deleting databases, as an examples of big data real time analysis. In this research, we used Twitter streaming data as an input, Apache Storm as a processing platform and D3.js as a visualization tool.

Finally, the designed dashboard was evaluated using Design of Experiment method and other statistical tests in various types of Apache Storm configurations and eventually it was proved that the dashboard is real time with an average response time for 1 minute and 30 seconds.

Keywords: Big data, visualization, real time dashboard

۱- مقدمه

می‌دانیم با پیشرفت سریع فناوری، حجم داده‌هایی که در این بستر منتقل و ذخیره می‌شوند، رو به فزونی گذاشته است. در چنین شرایطی است که واژه کلان‌داده پا به عرصه ظهور می‌گذارد. کلان‌داده با ویژگی‌های اصلی حجم بالا، سرعت بالا و تنوع در فرمت داده شناخته شده و در ابتدا جهت عمومی‌سازی مفهوم کلان‌داده، انجمن‌های مطالعاتی و رسانه‌ها به اصوات یکسان در ابتدای تعریف توجه کرده و از اصطلاح 3V، حجم (Volume)، سرعت (Velocity) و تنوع (Variety) استفاده می‌کنند. سایرین نیز از این روش استفاده کرده و Vهای دیگر نظیر ارزش (Value)، یا قابلیت تغییرپذیری (Variability) جهت بهبود و ارتقای معنا به تعریف بالا افزوده‌اند [1]. در سال‌های اخیر کلان‌داده با ویژگی‌هایی که برای آن برشمرده می‌شود، شناخته شده و تعریف می‌شود. سازمان مطالعاتی گارتنر می‌تواند گزینه معتبری برای شناسایی معانی رایج و مورد استفاده در این حوزه به‌شمار آید: کلان‌داده دارایی‌های اطلاعاتی با حجم، سرعت و تنوع بالا بوده که برای افزایش درک و قدرت تصمیم‌گیری، نیازمند پردازش اطلاعاتی از نوع نوآورانه و مؤثر است؛ همچنین به گزارش همین سازمان در سال ۲۰۱۶ کلان‌داده، مجموعه‌ای از فناوری‌های مختلف مدیریت داده که تجزیه و تحلیل‌های موارد کاربردی[1] چندگانه را پشتیبانی می‌کند، تعریف شده است. امروزه واژه کلان‌داده به یکی از واژه‌های کلیدی علم داده تبدیل شده است. منابع داده بسیار متنوع بوده و همچنین سرعت تولید داده‌ها با انواع فرمت داده رو به فزونی گذاشته است. تحلیل این حجم عظیم از داده‌ها که دارای تنوع بسیار بوده و به‌سرعت در حال افزایشند، ساختارهای سنتی تحلیل داده را با مشکلات بسیاری روبه‌رو ساخت. درنتیجه این مشکلات، پلتفرم‌ها، زیرساخت‌ها و ابزارهای نوینی برای تجمیع، پردازش، تحلیل و درنهایت بصری‌سازی کلان‌داده معرفی شده‌اند، که امروزه شاهد معرفی و توسعه هرچه بیشتر این ابزارها هستیم. یکی از مهم‌ترین دغدغه مدیران و تحلیل‌گران داده، تحلیل داده به‌صورت بلادرنگ است. جهان امروز به‌سرعت در حال تغییر است و لازمه وجود بقا در دنیای رقابتی امروز، گرفتن تصمیمات فوری بر مبنای داده‌های جریانی است.

باید توجه داشت یکی از منابع ارزشمند داده‌های جریانی، داده‌های تولیدشده کاربران شبکه‌های اجتماعی است. منابع داده‌ای شبکه‌های اجتماعی از آنجایی که سرچشمه گرفته از نظرات و تفکرات کاربران آنهاست، منابع بسیار غنی برای تحلیل محسوب می‌شوند. از میان شبکه‌های اجتماعی، شبکه اجتماعی توییتر نیز از آنجا که برای افراد

[1] Usecases

[DOI: 10.29252/jsdp.16.4.151]

به‌راحتی این امکان را فراهم می‌آورد که بتوانند علاوه‌بر شبکه آشنایان خود با اشخاص برجسته و کسانی که اطلاعات ارزشمندی را در اختیار آنان قرار می‌دهند، ارتباط برقرار کنند، در میان نسل‌های مختلف محبوبیت زیادی را کسب کرده است [2].

در این پژوهش تلاش بر آن بوده است که انواع ابزارها و زیرساخت‌های داده‌های جریانی کلان‌داده در حوزه‌های متفاوت بررسی شده و بر مبنای بررسی‌های انجام‌شده یک داشبورد بلادرنگ بر مبنای داده‌های جریانی شبکه اجتماعی توییتر طراحی شود.

به‌طورکلی مقاله پیش رو دو سؤال اساسی را دنبال می‌کند:

- نحوه طراحی و پیاده‌سازی یک داشبورد بلادرنگ بر مبنای شبکه‌های اجتماعی چگونه است؟
- کدام پیکره‌بندی‌های مختلف برای تحلیل و بصری‌سازی داشبورد بلادرنگ مناسبتر هستند؟

در ادامه و در بخش دوم، مفاهیم مرتبط با چیستی کلان‌داده و تحلیل‌های آن، بصری‌سازی و داشبوردهای بلادرنگ معرفی خواهند شد؛ سپس در بخش سوم پژوهش‌های انجام‌شده مرتبط با پژوهش حاضر بررسی خواهند شد. در بخش چهارم روش پژوهش را بررسی کرده و در بخش پنجم نحوه طراحی و ارزیابی داشبورد بلادرنگ ارائه خواهد شد و درنهایت در بخش پایانی نتایج جمع‌بندی شده و مورد بحث قرار خواهند گرفت.

۲- مبانی نظری

امروزه با توجه به پیشرفت فناوری و زیستن در عصر اطلاعات و ارتباطات، با داده‌هایی روبه‌رو هستیم که از نظر افزایش در ویژگی‌هایی چون سرعت تولید، حجم و تنوع، برایمان غیر قابل تصوراند. داده‌ها در حجم بی‌سابقه‌ای تولید و جمع‌آوری می‌شوند؛ و این امر منجر به ظهور مفهوم کلان‌داده شده است. ظرفیت بشر برای تولید داده تا کنون از زمان پیدایش فناوری اطلاعات در اوایل قرن نوزدهم تا این حد قدرتمند و گسترده نبوده است [3]. به گزارش IDC[1]، مقدار داده هر ساله دو برابر می‌شود و به حدود ۴۴ زتا بایت تا سال ۲۰۲۰ خواهد رسید. برای مثال، داده‌های حاصل از سامانه‌های تعبیه‌شده[2] از دو درصد در سال ۲۰۱۳ به ده درصد در سال ۲۰۲۰ خواهد رسید. تعریف علمی‌تر بیان می‌کند که کلان‌داده، داده‌هایی با اندازه‌ای فراتر از توانایی ابزارهای نرم‌افزاری رایج مورد استفاده برای جمع‌آوری، انتخاب و سازماندهی، مدیریت، و پردازش داده‌ها در مدت زمان سپری‌شده قابل قبول هستند [4]. در برخی از منابع، مانند [5] به جای ویژگی قابلیت تغییرپذیری از Veracity یا همان صحت داده‌ها برای تعریف کلان‌داده استفاده شده است. حال که به تعریفی دقیق از کلان‌داده دست یافته‌ایم به بررسی تأثیر ویژگی‌های کلان‌داده بر روی تحلیل داده‌ها و بصری‌سازی آنها خواهیم پرداخت.

تأثیرات کلان‌داده بر تحلیل اطلاعات را می‌توان در چهار بخش مختلف در نظر گرفت. بخش نخست استخراج و پردازش داده، بخش دوم تحلیل داده‌ها، بخش سوم ذخیره‌سازی آنها و درنهایت خود بصری‌سازی داده‌هاست.

پردازش داده: در مطالعات مختلف دسته‌بندی‌های متفاوتی از پردازش کلان‌داده ارائه شده است. مهم‌ترین نگاهی که به پردازش کلان‌داده در این بررسی وجود دارد نگاهی است که در مطالعات ابراهیم هاشم و همکاران [6] به آن اشاره شده است. پردازش کلان‌داده در این پژوهش به دو دسته تقسیم شده است. این دو نوع عبارتند از: دسته‌ای[3] و بلادرنگ[4]. در مطالعه آباواژی [7] این دو دسته پردازش که در هر راه حل جامع کلان‌داده در این روزها استاندارد هستند این‌گونه معرفی شده‌اند: پردازش دسته‌ای کلان‌داده مربوط به پردازش برون‌خط داده است. هادوپ یکی از محبوب‌ترین فناوری‌ها برای پردازش موازی فایل‌های حجیم در حالت دسته‌ای است. پردازش بلادرنگ به‌طورمعمول برای تحلیل داده‌های جریانی به‌صورت بلادرنگ‌بدون ذخیره‌سازی آن در دیسک به کار برده می‌شود. داده به‌صورت بلادرنگ تحلیل و پردازش می‌شود، همان‌طورکه از منابع مختلف جریان می‌یابد، تا یک بینش فوری به‌دست آید.

تحلیل داده: تحلیل داده‌ها بیش از هر چیزی متأثر از فرمت داده‌های گوناگونی است که امروزه با آنها روبه‌رو هستیم. در مطالعات بندر و تول [8] تحلیل داده‌ها متأثر از

[1] International Data Corporation

[2] embedded
[3] Batch
[4] Real time

[DOI: 10.29252/jsdp.16.4.151]

فرمتشان در رده‌های مختلفی تقسیم‌بندی شده‌اند، که عبارتند از: تجزیه و تحلیل داده ساختاریافته، تجزیه و تحلیل متن، تجزیه و تحلیل داده وب، تحلیل داده چندرسانه‌ای، تجزیه و تحلیل شبکه و داده‌های موبایل.

ذخیره‌سازی داده: رشد سریع داده، ظرفیت فناوری‌های ذخیره‌سازی فعلی را برای ذخیره‌سازی و مدیریت داده‌ها محدود کرده است [6]. اما، بیشتر سیستم‌های ذخیره‌سازی برای ذخیره‌سازی و مدیریت کلان‌داده دارای محدودیت بوده و نامناسب هستند. برای ذخیره‌سازی و مدیریت مجموعه داده‌های حجیم به یک معماری ذخیره‌سازی که بتواند به‌شیوه‌ای بسیار کارامد دسترس‌پذیری و قابلیت اطمینان را تأمین نماید نیاز است [6]. درنتیجه عصر کلان‌داده، محدودیت‌های بسیاری مربوط به سامانه‌های مدیریت پایگاه داده‌های رابطه‌ای به خود جلب کرده است، به‌خصوص زمانی‌که فرمت‌های داده‌های پیچیده را مدیریت می‌کنند. با رشد دسترس‌پذیری داده و افزایش عدم تجانس آن، درحقیقت، مقیاس‌پذیری و انعطاف‌پذیری به نگرانی‌های مهم و اصلی تبدیل شده‌اند. کاربردهای علمی فعلی با مقادیر بالای دادهٔ نیمه‌ساختاریافته و غیرساختاریافته از قبیل آرایه‌های چندبعدی، گراف‌ها و شبکه‌های نامرتب[1] سر و کار دارند، که نمی‌توانند به شیوه‌ای رابطه‌ای نشان داده شوند؛ بنابراین، از الزامات اصلی نسل بعدی پایگاه داده‌ها، قابلیت خواندن، ویرایش، به‌روزرسانی منابع داده‌های بدون ساختار، نسخه‌بندی داده‌ها و پیگیری سرچشمه داده است، بدون آن‌که نیاز باشد نسخه‌ای پشتیبان از داده‌ها ایجاد شود. پایگاه داده‌های NoSQL به‌صورت گسترده‌ای برای غلبه بر عدم انعطاف‌پذیری پایگاه داده‌های رابطه‌ای با توجه به داده‌های ناهمگن استفاده می‌شوند و به ارائه پشتیبانی بهبودیافته برای پرس‌وجوهای[2] توزیع‌شده و ذخیرهٔ یکپارچه می‌پردازند [9]. پایگاه داده‌های NoSQL می‌تواند با توجه به تقسیم بندی‌های ارائه‌شده در پژوهش‌های تئودوریکا و همکاران [10]، کتل [11]، هچت و همکاران [12] و لیویت [13] که هر کدام طرح خاص برای دادهٔ ذخیره‌شده تجویز می‌کنند، به چندین دسته تقسیم شوند: پایگاه داده‌های مبتنی بر مستندات[3]، مبتنی بر ستون[4]، مبتنی بر گراف[5] و کلید-ارزش[6] [14].

بصری‌سازی: همان‌طور که پیش‌تر گفتیم، در عصر کلان‌داده، مقادیر زیادی داده به‌طور پیوسته برای اهداف متنوعی یافت می‌شوند. محاسبات پیشرفته[7] و فناوری‌های تصویربرداری و حس‌گر[8] پژوهش‌گران را قادر می‌سازند تا به مطالعه پدیده‌های طبیعی و فیزیکی در دقت بی‌سابقه‌ای بپردازند، که این امر خود رشد انفجاری داده را موجب می‌شود. بصری‌سازی این حجم دادهٔ روبه‌رشد چه به شکل ایستا و یا پویا چالشی بزرگ محسوب می‌شود. بسیاری از رویکردها و ابزارهای سنتی بصری‌سازی در مقیاس بزرگ پاسخ‌گو نخواهند بود. بصری‌سازی داده برای درک داده توسط افراد، به شیوه‌ای گرافیکی بسیار کارآمد است؛ اما، در عصر کلان داده، داده به‌صورت پیوسته بزرگ و بزرگ‌تر می‌شود. نمایش معنی‌دار و با ارزش داده با حجم بالا بسیار دشوار است. بصری‌سازی داده سنتی برای هدایت و مدیریت کلان‌داده ناکافی و نامناسب است. برای مثال، بسیاری از مجموعه‌داده‌ها برای جای‌گذاری در حافظه بسیار حجیم بوده و شاید در سراسر یک خوشه توزیع شوند [15]. بصری‌سازی دادهٔ جدید باید شیوه‌های بهتری را برای پردازش، تحلیل و بصری‌سازی مقادیر بالای داده‌های پیچیده ارائه دهد.

درهمین‌اواخر، موضوع تحلیل‌های بلادرنگ داده‌های پیچیده کلان نیز باید مورد توجه قرار گیرند. همان‌گونه که رشد داده‌ها به‌طور پیوسته بیشتر و بیشتر می‌شود، راه‌های سنتی نمایش داده‌ها به محدودیت‌های بیشتری برخورد می‌کند [16]. برای مثال، چگونه تمامی داده‌های قابل تصویرشدن با حجم بالا را بر روی صفحه‌های نمایش محدود نمایش داد (بررسی داده‌های با حجم زتا بایت به جای گیگا بایت). علاوه‌بر آن، پرس‌وجوی[9] روان و بدون اختلال در مجموعه‌داده‌های کلان با مشکل مواجه می‌شود.

همچنین نمایش داده‌های حجیم به‌صورت بلادرنگ بسیار اهمیت دارد. این محدودیت‌ها چالش‌هایی نظیر

[1] Irregular meshes
[2] Queries

[3] Document-oriented
[4] Column-oriented or wide column
[5] Graph-oriented
[6] Key-value
[7] Advanced computing
[8] sensing
[9] Query

[DOI: 10.29252/jsdp.16.4.151]

مقیاس پذیری ادراکی[1]، مقیاس پذیری بلادرنگ[2] و مقیاس‌پذیری تعاملی[3] به‌وجود آورده است [3]. وقتی صحبت از کلان‌داده است، نحوه و چگونگی نمایش آن می‌تواند به انتقال اطلاعات کمک کند؛ اما نمایش آن به چیزی فراتر از فقط ظاهری زیباداشتن نیاز دارد. داشبورد کلان‌داده باید به‌درستی کار کند، قابلیت نمایش ابعاد چندگانه را داشته باشد و در انتقال اطلاعات به کاربر نهایی مؤثر واقع شود. بصری‌سازی کلان‌داده همچنین فرصت‌هایی برای راه‌های بهتر نمایش برای بصری‌سازی کلان‌داده مانند داده‌کاهی[4] و کاهش زمان تأخیر[5] و ... به وجود می‌آورد [17].

۳- مبانی تجربی (پژوهش‌های پیشین)

حال که تأثیرات کلان‌داده مورد بررسی قرار گرفتند، لازم است گفته شود هدف اجرای این پژوهش طراحی و پیاده‌سازی داشبوردی بلادرنگ متأثر از پردازش‌های جریانی که نمونه‌ای از داده‌های متنی توییتر را با روش ابر واژه بصری‌سازی می‌کند، بوده است.

به‌دلیل نیاز علمی و کاربردی در حوزه کلان‌داده امروزه پژوهش‌ها و مطالعات بسیاری در زمینه‌های مختلف انجام می‌شود. حوزه‌های پژوهشی در کلان‌داده بسیار گسترده بوده و هر کدام به‌صورت جزئی، بخشی از این حوزه را مانند انواع پردازش، انواع ذخیره‌سازی و ... مورد بررسی قرار می‌دهند. گفتنی است در کشور ما، پژوهش‌های صورت‌گرفته نسبت به پژوهش‌های بین‌المللی اندک بوده ولی سیری صعودی در پیش گرفته است. یکی از زمینه‌های مهم پژوهشی تحلیل صحیح کلان‌داده و نمایش آن به‌نحوی قابل فهم برای کاربران نهایی است. در جدول (۱) خلاصه‌ای از آنچه که در پژوهش‌های بین‌المللی به موضوع مطرح پرداخته است، آورده شده است.

[1] Perceptual scalability
[2] Real time scalability
[3] Interactive scalability
[4] Data reduction
[5] Reducing Latency

(جدول-۱): پژوهش‌های مرتبط با بصری‌سازی کلان‌داده

(Table-1): Research related to massive visualization of data

پژوهش‌گر	سال	عنوان	یافته‌ها
گرول و همکاران [18]	۲۰۱۰	هوشمندی اجتماعی چندمنظوره در یک سامانه داشبورد بلادرنگ	هدف از انجام این پژوهش طراحی و پیاده‌سازی داشبوردی بوده است که بتواند با استفاده از داده‌های کلان شبکه‌های اجتماعی، تارنماهای متفاوت و استخراج اطلاعات از کامنت‌های افراد، فهرستی به‌طورتقریبی به هنگام از آهنگ‌های محبوب در اختیار کاربران بی بی سی قرار دهد. در پایان اظهار می‌شود پیاده‌سازی چنین سامانه‌های چندحالته که در آنها منابع داده متفاوت بوده و دائماً در حال تغییر هستند، و همچنین داده‌ها به‌طورمعمول غیر ساخت یافته‌اند، نسبت به سامانه‌های ایستا بسیار دشوار بوده و نیازمند به کارگیری روش‌ها و ابزار پیچیده‌تری هستند.
بنجامین فلش[19]	۲۰۱۴	طراحی، توسعه و ارزیابی یک داشبورد تحلیلی کلان‌داده	هدف از انجام این پژوهش طراحی، پیاده‌سازی و ارزیابی داشبوردهای کلان داده بوده است. در این پژوهش با استفاده از ابزار SODATO ابتدا داده‌های صفحه‌های فیسبوک مربوط به یازده برند پوشاک در اروپا و آمریکا در طی یک بازه زمانی استخراج و سپس اخبار مربوط به تولیدی‌های پوشاک در بنگلادش در همین بازه زمانی استخراج می‌شود. بعد از پایش داده‌ها این دو موضوع در کنار هم در یک ابزار بصری‌سازی متن باز که برای نمایش کلان‌داده به کار می‌رود، نشان داده شده‌اند. این داشبورد درنهایت برای ارزیابی از نظر سهولت استفاده، کاربرپسندبودن و ... در اختیار کاربران قرار می‌گیرد و مورد تأیید واقع می‌شود.

[DOI: 10.29252/jsdp.16.4.151]

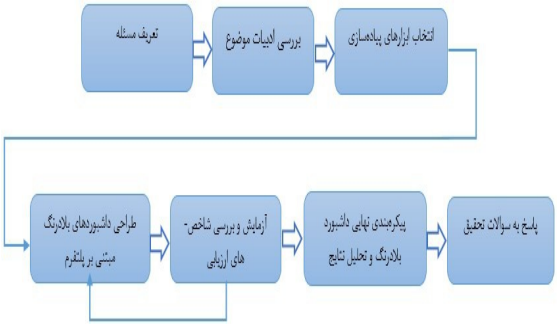
پژوهش‌گر	سال	عنوان	یافته‌ها
هوبر و همکاران [20]	۲۰۱۵	تجزیه و تحلیل بصری توییتر: کشف احساسات در حال تغییر و کشف زمینه‌های نو ظهور در توییت رویداد ورزشی	در این پژوهش یک سامانه نرم‌افزاری به نام Vista تولید شده است. هدف از توسعه این سامانه آنالیز بصری داده‌های متنی توییتر است. داده‌های توییتر توسط Twitter API استخراج شده، سپس در یک پایگاه داده ذخیره می‌شوند، سپس با استفاده از الگوریتم‌های طبقه‌بندی و با استفاده از sentiment140،پیام‌های متنی توییتر در سه طبقه مثبت،منفی و خنثی قرار گرفتند. برای بصری‌سازی داده‌ها و توسعه داشبورد از D3 استفاده شده است؛ سپس سامانه تولیدشده در یک مطالعه موردی، یعنی مجموعه‌داده‌ای شامل ۴۰۹۰۰۰ توییت مربوط به مسابقات دوچرخه‌سواری توردوفرانس این سامانه آزمایش شده و بهره‌برداری شده است.
ویلر و همکاران [21]	۲۰۱۵	نظارت بر وضعیت مناطق شهری با استفاده از جریان داده‌ای رسانه‌های اجتماعی	رشد شبکه‌های اجتماعی منجر به تولید حجم انبوهی از داده‌ها توسط کاربران شده است. در این مقاله گفته می‌شود هدف استفاده از کاربران شبکه‌های اجتماعی به‌عنوان "حس‌گر اجتماعی" به‌منظور افزایش آگاهی موقعیتی در مورد مناطق شهری است. برای این منظور توییت‌های تولیدشده توسط کاربران توییتر از نظر رویدادها و موضوعات در بازه‌های زمانی و مکانی مختلف استخراج شده و پس از پردازش در قالب ساعت و tag cloud به نمایش در آمده‌اند. برای این منظور مجموعه‌داده مطالعه موردی‌های انجام‌گرفته از Twitter Streaming API استخراج شده است. برای پردازش سامانه مدیریت جریان داده‌ها به نام Niagarino استفاده شده است. درنهایت بصری‌سازی داده‌ها توسط tag cloud در قاب موجود worderam انجام گرفته شده است.
بابک یادرنجی اقدم |22|	۲۰۱۶	توسعه یک چارچوب تحلیل داده‌ای بلادرنگ برای داده‌های جریانی توییتر	در این پژوهش چارچوبی برای تحلیل بلادرنگ داده‌های توییتر ارائه شده است. این چارچوب برای جمع‌آوری، فیلتر و تجزیه و تحلیل جریان داده‌ها طراحی شده است. چارچوب شامل سه مرحله اصلی است: استخراج داده‌ها، پردازش داده‌های جریانی و بصری‌سازی داده‌ها. استخراج داده‌ها در مرحله نخست توسط کافکا انجام می‌شود. تجزیه و تحلیل داده‌ها در مراحل بعدی، با استفاده از Apache Spark و با بهره‌برداری از الگوریتم های یادگیری ماشینی انجام شده است؛ سپس از چارچوب ارائه‌شده در این پژوهش در یک مطالعه موردی در مورد زمین‌لرزه در ژاپن در نوامبر ۲۰۱۶ استفاده شده است. در طی مطالعات انجام‌شده داده‌های وارده (توییت‌ها) در طی یک بازه هشت‌ساعته مورد تحلیل قرار گرفتند. توییت‌های استخراج‌شده از نظر منشأ و تعداد پس از پخش خبر زلزله مورد بررسی قرار گرفتند. در طی این مطالعه پژوهش‌گر به بینش بسیار خوبی در مورد نحوه واکنش مردم به‌طور خاص به حوادث غم انگیز یا خطرناک در مکان‌های مشخص دست می‌یابد.

[DOI: 10.29252/jsdp.16.4.151]

از میان مطالعات یادشده، دو پژوهشی که بیش از سایرین این مقاله را متأثر ساخته است، عبارتند از مطالعات بنجامین فلش [19] و بابک یادرنجی اقدم [22,23]. در مطالعات بنجامین فلش برای ورودی از داده‌های شبکه اجتماعی فیسبوک استفاده و از پردازش دسته‌ای بهره گرفته و در نهایت ساخت داشبورد با D3.js صورت گرفته است. در پژوهش‌های بابک یادرنجی اقدم همانند این پژوهش از داده‌های جریانی توییتر استفاده و از بستر پردازشی Spark Streaming برای پردازش داده‌ها بهره گرفته شده است.

۴- روش پژوهش

برای انجام پژوهش پس از تعریف مسأله، به‌مرور و بررسی ادبیات موضوع پرداخته شد. در زمینه ادبیات موضوع، کلان‌داده و موضوعات مرتبط با آن بسیار گسترده بوده و تمامی جنبه‌های مختلف آن مورد علاقه دانشمندان و پژوهش‌گران این حوزه است؛ اما گفتنی است به تحلیل‌های بلادرنگ و انتخاب ترکیبی از ابزارهای مناسب برای پیاده‌سازی این تحلیل‌ها در محیط‌های دانشگاهی و پژوهشی کمتر پرداخته شده است. با توجه به تغییرات دنیای امروز و افزایش سرعت تولید داده‌هایی که می‌توانند برای تصمیم‌گیری مورد استفاده مدیران قرار گیرند، لازم است به این زمینه از پژوهش‌های کلان‌داده بیشتر توجه شود. بنابراین از آنجا که این پژوهش در نظر دارد داشبورد بلادرنگ کلان‌داده را مورد بررسی قرار دهد و ویژگی‌های آن را با داشبوردهای سنتی مقایسه کرده و درنهایت اقدام به طراحی و پیاده‌سازی آن کند، می‌توان آن را از نظر هدف، کاربردی دانست. همچنین با توجه به استفاده از منابع مختلف و داده‌های شبکه‌های اجتماعی، از لحاظ نوع داده‌ها، کمی و کیفی بوده و برای تجزیه و تحلیل داده‌ها و نتایج از روش آزمایشی بهره جسته است.



(شکل-۱): فرآیند انجام پژوهش
(Figure-1): Research Process

(جدول-۲): اجزا و ویژگی‌های ابزارهای معماری داشبورد بلادرنگ
(Table-2): Components and features of real-time dashboard architecture

اجزای ساخت	ابزار استفاده شده
ورودی	داده‌های جریانی توییتر
مجازی‌سازی	Virtualbox Vagrant
ماشین مجازی	Operating System: Ubuntu (32-bit) Base Memory: 2048 MB
استخراج داده	Twitter Streaming API
پلتفرم پردازشی	آپاچی استورم 0.9.2-Incubating
تکنیک‌های مورد استفاده	فیلترینگ، تناوب
زبان برنامه‌نویسی	Java
ابزار بصری‌سازی	D3.js

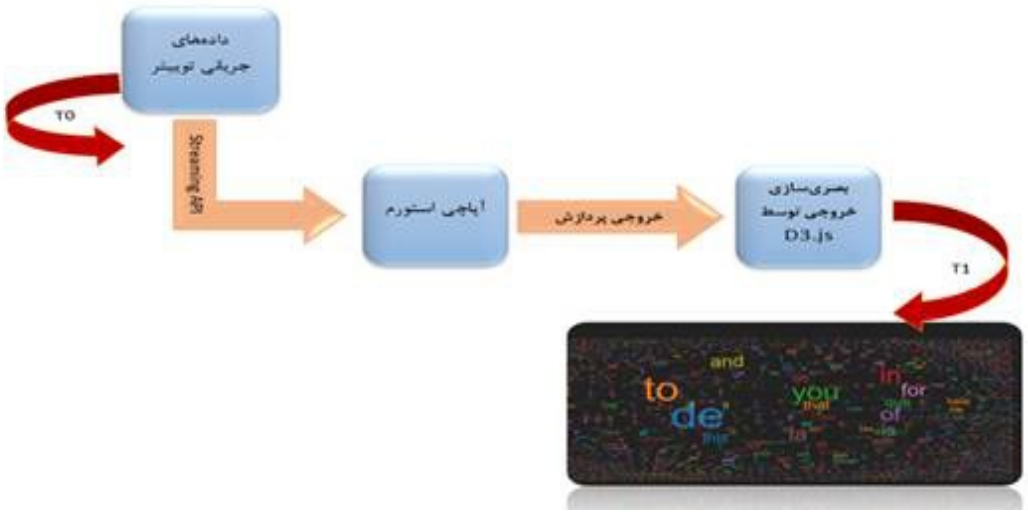
پس از بررسی ادبیات موضوع، به مطالعه ابزارهای مختلف در این زمینه پرداخته شد. درنهایت از بین ابزارهای متن باز و تجاری مختلف درنهایت همان‌طور که اشاره شد، از ترکیب آپاچی استورم و D3.js بهره گرفته شد. دلیل انتخاب آپاچی استورم آن است که این پلتفرم پردازشی برای پردازش داده‌های بلادرنگ طراحی شده و دلیل انتخاب D3.js، توانایی آن در ارائه انواع بصری‌سازی‌ها است؛ سپس طراحی فنی داشبورد انجام گرفته و داشبورد بلادرنگ پیاده‌سازی‌شده با روش‌های مختلف مورد ارزیابی قرار گرفت؛ سپس نتایج حاصل از ارزیابی‌ها ارائه و سؤالات پژوهش پاسخ داده می‌شوند. گام‌های پژوهش به‌صورت طرح‌واره در شکل (۱) و اجزای ساخت داشبود در جدول (۲) نشان داده شده‌اند.

۵- طراحی و ارزیابی

۵-۱- طراحی

برای ساخت یک داشبورد بلادرنگ ابتدا اجزای مورد نیاز شناسایی شدند. اجزای ساخت یک داشبورد بلادرنگ عبارتند از: داده‌های جریانی ورودی، ابزار پردازش داده و درنهایت بصری‌سازی آن. برای ساخت داشبورد بلادرنگ این پژوهش پس از بررسی‌های صورت‌گرفته، برای داده ورودی از داده‌های جریانی استفاده و سپس برای ابزار پردازش و بصری‌سازی به‌ترتیب آپاچی استورم و D3.js انتخاب شدند. معماری داشبورد به‌صورت طرح‌واره در شکل (۲) نشان داده شده است.

[DOI: 10.29252/jsdp.16.4.151]



(شکل-۲): معماری طرح‌واره داشبورد

(Figure-2): Schematic architecture of the dashboard

داده‌های جریانی توییتر در زمان T0 توسط Streaming API استخراج شده و پس از پردازش توسط آپاچی استورم و بصری‌سازی با D3.js در زمان T1 در داشبورد نمایش داده می‌شوند. برای طراحی محتوای داشبورد از سه سناریو بهره برده می‌شود.

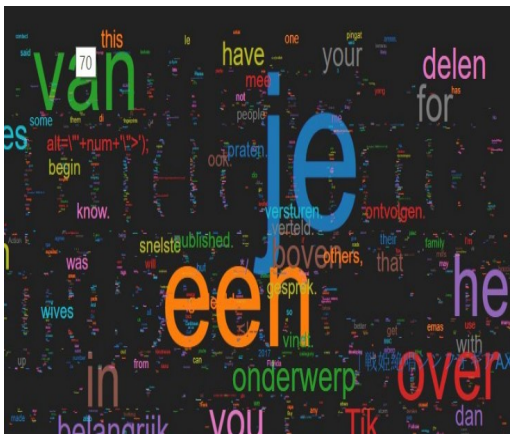
سناریوی نخست: شمارش واژگان توییت‌های جریان‌یافته و نمایش آنها با روش ابر واژه

سناریوی دوم: شمارش واژگان URLهای توییت‌های کاربران

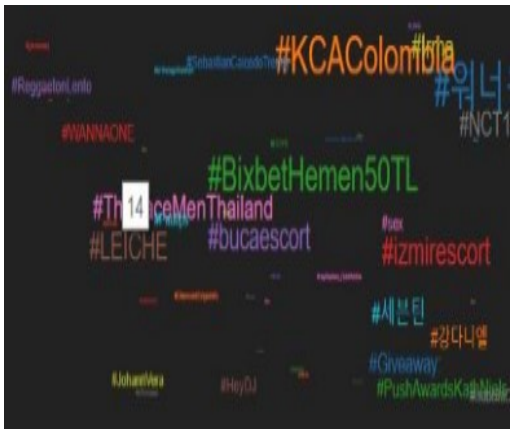
سناریوی سوم: یافتن بیشترین هشتگ‌ها و نمایش آنها با روش ابر واژه

برای هر کدام از سناریوها خروجی به شکل زیر است:

داشبورد سناریوی نخست: شمارش واژگان توییت‌های جریان‌یافته و نمایش آنها با روش ابر واژه

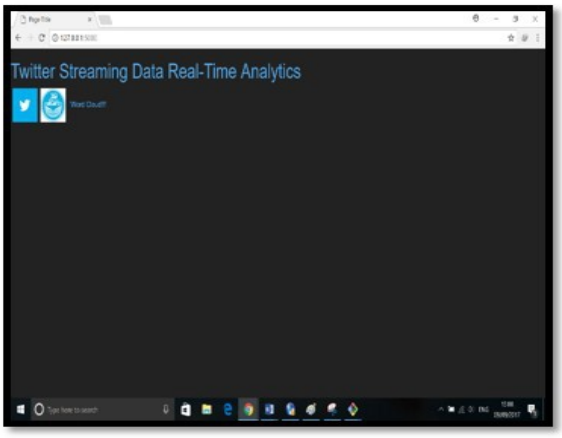


داشبورد سناریوی دوم: شمارش واژگان URLهای توییت‌های کاربران



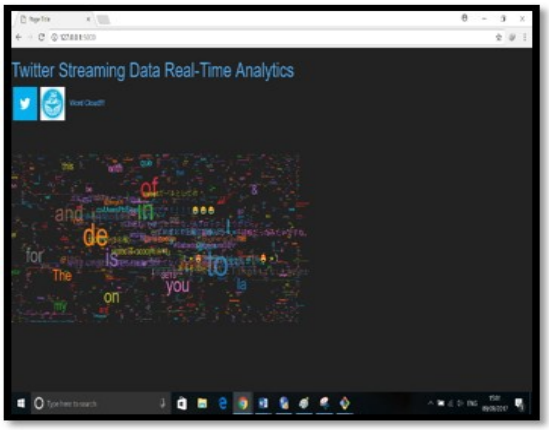
سناریوی سوم: یافتن بیشترین هشتگ‌ها و نمایش آنها با روش ابر واژه

[DOI: 10.29252/jsdp.16.4.151]

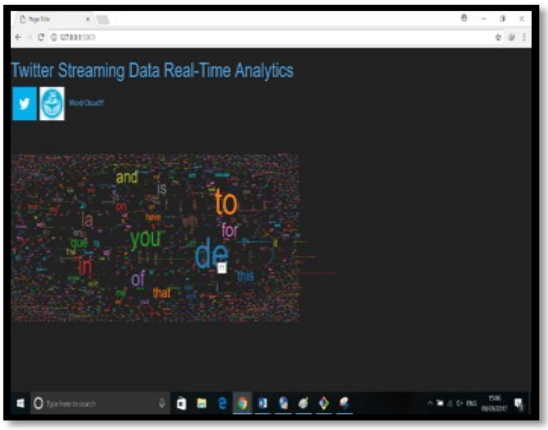


شکل آ. قبل از اجرای توپولوژی

شکل ب. ۲۰ ثانیه پس از اجرا



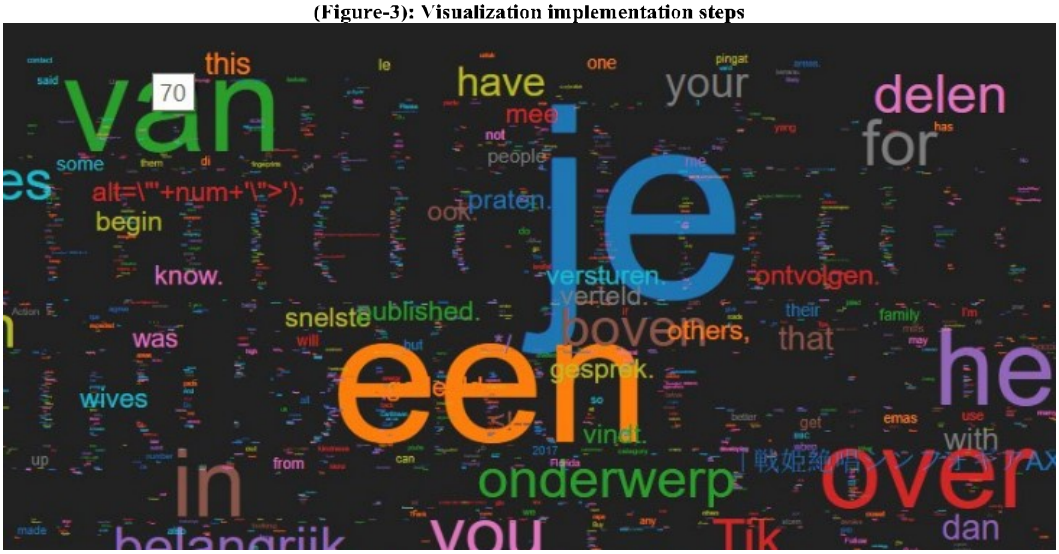
شکل پ. ۵۰ ثانیه پس از اجرا



شکل ت. اتمام اجرای توپولوژی

(شکل-۳): مراحل اجرای بصری‌سازی

(Figure-3): Visualization implementation steps



(شکل-۴): نمایی نزدیک از داشبورد پارس‌شدن واژگان توییت‌ها

(Figure-4): A close view of the parsed Tweets dashboard

[DOI: 10.29252/jsdp.16.4.151]

یکی از مهم‌ترین بخش‌های پیاده‌سازی داشبورد بلادرنگ توییتر در این پژوهش نحوه پیکره‌بندی آپاچی استورم برای پردازش داده‌ها است. برای توضیح نحوه پیکره‌بندی، لازم است ابتدا اشاره‌ای مختصر به مفاهیم آن داشته باشیم. اجزای پلتفرم پردازشی آپاچی استورم عبارتند از اسپات[1] و بولت‌ها[2]. اسپات وظیفه دریافت جریان ورودی را بر عهده داشته و بولت‌ها الگوریتم‌های پردازشی را بر روی داده‌ها اجرا می‌کنند. مفهوم دیگری که لازم است در این بخش به آن اشاره شود، گروه‌بندی جریان[3] است. گروه‌بندی جریان، نحوه اتصال بین وظایف[4] اسپات و بولت‌ها را مشخص می‌سازد. در این پژوهش دو نوع گروه‌بندی جریان که مورد ارزیابی قرار گرفتند، عبارتند از: All grouping و Shuffle grouping.

در All grouping جریان در سراسر تمام وظایف بولت‌ها تکثیر می شود؛ اما درShuffle grouping، جریان داده به طور تصادفی در سراسر وظایف بولت توزیع شده، به‌طوری‌که تضمین شده است به هر یک از بولت‌ها تعداد یکسانی از تاپل‌ها[5] وارد شود؛ درنهایت اسپات و بولت‌ها و نحوه اتصال آنها را پیکره‌بندی می‌نامند. به‌عبارت دیگر پیکره‌بندی همان معماری پلتفرم پردازشی ما یعنی آپاچی استورم است. در سناریوی نخست و دوم، نمونه‌ای از داده‌های توییتر توسط Streaming API استخراج شده و برای پردازش توسط آپاچی استورم توسط اسپات‌ها در پیکره‌بندی جریان می‌یابند. برای اجرای الگوریتم شمارش واژگان (Word Count) توییت‌های جریانی توسط ایستگاه ابتدایی خوانده شده؛ سپس شمارش تعداد واژگان خوانده‌شده در ایستگاه بعدی صورت گرفته و درنهایت وارد ایستگاه گزارش شده و نتایج تحلیل نهایی توسط D3.js مصور می‌شوند.

در سناریوی سوم نیز همانند روش قبل نمونه‌ای از داده‌های توییتر استخراج شده و برای پردازش توسط آپاچی استورم توسط اسپات‌ها در پیکره‌بندی جریان می‌یابند. تفاوت پیکره‌بندی در به‌کارگیری بولت‌های متفاوت است. در این سناریو تنها شمارش هشتگ‌ها هدف نهایی نیست و بلکه برگرداندن بالاترین هشتگ‌های (TopN) توییت شده است. درنتیجه بولت پردازشی ranking برای نمایش بیشترین هشتگ‌ها به کار گرفته شده است.

در شکل‌های (۳ و ۴) روند بصری‌سازی نشان داده شده است. همان‌طورکه در شکل (۴) مشخص است، جملات هر توییت به واژگان سازنده‌اش شکسته شده و با قرارگرفتن بر روی هر واژه، تعداد آن نمایش داده می‌شود.

۵-۲- ارزیابی

پس از مرور ادبیات مانند مطالعات بابک یادرنجی اقدم [22,23] و کوردووا [24] این نتیجه کسب شده که فرایند ارزیابی داشبوردهای بلادرنگ بر اساس میزان زمان طی‌شده از تولید محتوا تا دریافت پاسخ انجام می‌شود. در نتیجه هدف ارزیابی در این پژوهش نیز بلادرنگ‌بودن داشبورد طراحی‌شده تعریف شد که برای سنجش آن از شاخص میزان زمان تأخیر برای پاسخ استفاده شد. در این پژوهش برای ارزیابی داشبورد پیاده‌سازی شده از دو روش بهره برده می‌شود:

- سنجش میزان زمان صرف‌شده از تولید محتوا در توییتر تا نمایش در داشبورد.
- سنجش میزان زمان صرف‌شده از اجرای پیکره‌بندی تا نمایش داده در داشبورد.

روش ۱. سنجش میزان زمان صرف‌شده از تولید محتوا در توییتر تا نمایش در داشبورد.

در روش نخست در حساب کاربری گروه پژوهش محتوایی مشخص توییت می‌شود؛ سپس ورودی محیط پردازشی را به‌گونه‌ای فیلتر خواهیم کرد تا تنها اطلاعات حساب کاربری یادشده بازگردانده شود؛ پس از اجرای پیکره‌بندی زمان لازم، برگردانده شدن محتوای حساب کاربری سنجیده خواهد شد. پس از بررسی‌ها و اجرای چندین باره پیکره‌بندی، متوسط بازه زمانی یک دقیقه و سی ثانیه به‌دست می‌آید. از آنجا که تأخیر در ترافیک شبکه کاربر و ترافیک سرور را باید در نظر داشت زمان به‌دست‌آمده، زمان قابل قبولی خواهد بود.

روش۲. سنجش میزان زمان صرف شده از اجرای پیکره‌بندی تا نمایش داده در داشبورد.

روش دوم برای ارزیابی و بررسی دو فرضیه زیر طراحی شده است.

[1] Spout
[2] Bolt
[3] Stream grouping
[4] Tasks
[5] Tuples

[DOI: 10.29252/jsdp.16.4.151]

H0: تغییرات پیکره‌بندی آپاچی استورم بر زمان تأخیر مؤثر نمی‌باشد.

H1: تغییرات پیکره‌بندی آپاچی استورم بر زمان تأخیر مؤثر می‌باشد.

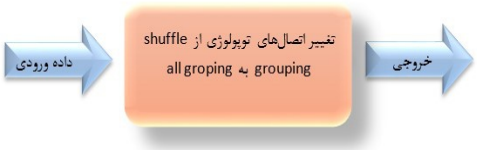
برای تحقق روش دوم و بررسی فرضیه‌های بالا از مدل ارزیابی طراحی آزمایش‌ها[1] استفاده خواهیم کرد.

پس از استخراج نتایج دو آزمایش، داده‌های هر دو آزمایش برای انجام آزمون اسمیرنف-کولموگروف وارد نرم‌افزار spss statistics شده و با انجام تحلیل درمی‌یابیم، توزیع نمونه نرمال است. ابتدا آزمایش با پیکره‌بندی مطرح در ابتدای مقاله برای هر سه سناریو به تعداد ده بار انجام شده و نتایج آن در جدول (۳) نمایش داده شده است؛ سپس پیکره‌بندی آپاچی استورم را تغییر داده و پیکره‌بندی را برای هر سناریو ۱۰ بار اجرا می‌کنیم. نحوه تغییر پیکره‌بندی آپاچی استورم در شکل (۵) آمده است.

(جدول-۳): آزمون اسمیرنف-کولموگروف

(Table-3): Smirnov-Kolmogorov test

		VAR00001	VAR00002
N		30	30
Normal Parameters[a,b]	Mean	23.0573	30.74
	Std. Deviation	3.80498	6.832
Most Extreme Differences	Absolute	.155	.295
	Positive	.155	.295
	Negative	-.115	-.231
Test Statistic		.155	.295
Asymp. Sig. (2-tailed)		.063[c]	.000[c]



(شکل-۵): تغییر در پیکره‌بندی آپاچی استورم

(Figure-5): Changes in Apache Storm configuration

نتایج حاصل بعد از اجرای پیکره‌بندی با پیکره‌بندی تغییریافته نیز در جدول (۵) آمده است.

(جدول-۴): پیکره‌بندی با اتصال shuffle

(Table-4): Topology with shuffle grouping

Word Count	Time(s)	URLs Parsing	Time(s)	TopN	Time(s)
1	27.24	1	18.87	1	**21.76**
2	29.36	2	21.27	2	**20.89**
3	23.71	3	19.15	3	**22.43**
4	25.97	4	18.63	4	**27.36**
5	26.44	5	22.53	5	**21.83**
6	30.17	6	21.72	6	**18.22**
7	29.17	7	19.29	7	**17.96**
8	28.19	8	18.95	8	**21.56**
9	26.97	9	23.46	9	**18.74**
10	27.32	10	22.34	10	**20.22**
Avg	**27,464**	**Avg**	**20.621**	**Avg**	**21.097**

(جدول-۵): پیکره‌بندی با اتصال all

(Table-5): Topology with all grouping

Word Count	Time(s)	URLs Parsing	Time(s)	TopN	Time(s)
1	64.28	1	25.10	1	**31.26**
2	26.51	2	27.42	2	**32.02**
3	33.16	3	22.85	3	**29.42**
4	29.42	4	24.17	4	**32.23**
5	26.44	5	24.32	5	**35.12**
6	33.03	6	26.14	6	**29.53**
7	31.22	7	27.13	7	**27.76**
8	27.09	8	25.67	8	**28.08**
9	26.46	9	24.26	9	**31.09**
10	31.15	10	25.02	10	**27.44**
Avg	**32.848**	**Avg**	**25.208**	**Avg**	**30.405**

پس از استخراج نتایج بالا، اطلاعات به‌دست‌آمده را با آزمون فرض آماری T در نرم‌افزار SPSS مورد تحلیل قرار داده و نتایج حاصل نشان می‌دهد که فرضیه H0 مبنی بر عدم تأثیر تغییرات پیکره‌بندی بر زمان تأخیر تأیید شده و فرضیه H1 رد می‌شود. البته باید در نظر داشت این آزمایش تنها زمان پاسخ را بررسی کرده و سایر ابعاد تحلیل مانند صحت‌سنجی نتایج تحلیل موضوعی خارج از بحث بوده است.

به‌اختصار نتایج حاصل از آزمایش در جدول (۶) آمده‌اند.

(جدول-۶): خلاصه نتایج حاصل از ارزیابی

(Table-6): Summary of the results of the evaluation

	آزمایش نخست	آزمایش دوم
تشریح آزمایش	انجام آزمایش با پیکره‌بندی استورم با اتصال‌های shuffle grouping	انجام آزمایش با پیکره بندی استورم با اتصال‌های all grouping
شاخص ارزیابی	زمان تاخیر ارزیابی	زمان تاخیر ارزیابی
عوامل ثابت آزمایش	Streaming API ابزارهای یکسان ماشین مجازی و سیستم یکسان	Streaming API ابزارهای یکسان ماشین مجازی و سیستم یکسان
عوامل متغیر آزمایش	پیکره‌بندی استورم با shuffle grouping	پیکره‌بندی استورم با all grouping
متغیرهای غیرقابل کنترل (تصادفی)	ترافیک سرور توییتر ترافیک شبکه کاربر	ترافیک سرور توییتر ترافیک شبکه کاربر
نتایج	Avg(Latency) = 23.06	Avg(Latency) = 29.487

[1] Design of experiments (DoE)

[DOI: 10.29252/jsdp.16.4.151]

توجه به این نکته حائز اهمیت است که برای پیاده‌سازی راه حل‌های کلان‌داده انتخاب ابزار مناسب با توجه به هدف طراحی از مهم‌ترین گام‌های پیاده‌سازی محسوب می‌شود؛ زیرا با توجه به پیدایش واژه کلان‌داده و درنظرداشتن ویژگی‌های مربوط به آن، امروزه شاهد افزایش نیاز هر چه بیشتر به تحلیل و استخراج دانش از کلان‌داده و درنتیجه متناسب با افزایش نیاز به تحلیل کلان‌داده، هم‌زمان شاهد رشد روزافزون ابزارها و راه حل‌های پشتیبان تجزیه و تحلیل‌های کلان‌داده هستیم. در این پژوهش نیز چون هدف پیاده‌سازی یک داشبورد بلادرنگ بود، آپاچی استورم که قابلیت پشتیبانی از پردازش‌های بلادرنگ را دارد انتخاب شد. در اینجا لازم است به این نکته اشاره شود که در تحلیل‌های بلادرنگ از آنجا که داده، بلافاصله پس از تولید وارد گام پردازش و تحلیل می‌شود، ابزار ذخیره‌سازی داده برای ذخیره‌سازی داده پس از تولید مورد نیاز نخواهند بود. گفتنی در پژوهش‌های پیشین مانند پژوهش بابک یادرنجی اقدم [22,23]، برای ارائه یک راه حل بلادرنگ از پلتفرم پردازشی Spark Streaming استفاده شده و در آن پژوهش نیز بلادرنگ‌بودن تحلیل اثبات شده است؛ اما باید به این نکته توجه داشت که در آن پژوهش به ابزار بصری‌سازی به خصوصی اشاره نشده و هدف پژوهش نبوده است. در پژوهش بنجامین فلش [19] نیز یک راه حل مبتنی بر کلان‌داده ارائه شده با این تفاوت که در راه حل ارائه‌شده توسط وی بلادرنگ‌بودن داشبورد مطرح نبوده و از داده‌های گذشته‌نگر[1] بهره گرفته شده است.

۶- مباحثه

نکته دیگری که در این بخش گفتنی است، آن است که از آنجا که امروزه شبکه‌های اجتماعی نقشی پر رنگ در زندگی افراد داشته و بیان‌گر اندیشه‌ها و نیازهای آنها در هر زمینه‌ای هستند، بنابراین اطلاعات آنها از منابع غنی برای استخراج دانش و بینش محسوب می‌شوند؛ درنتیجه صاحبان صنایع باید توجه داشته باشند. استخراج اطلاعات به‌صورت بلادرنگ از شبکه‌های اجتماعی می‌تواند آنها را در دست‌یابی به مزیت رقابتی بسیار یاری رسانده و جهت حرکت به‌سوی آینده را روشن ساخته و به آنها در جهت انجام برنامه‌ریزی‌های عملیاتی و استراتژیک کمک کند؛ درنتیجه وجود چنین نیازی است که در این پژوهش از داده‌های شبکه اجتماعی توییتر برای نمایش محتوای داشبورد استفاده شده است؛ درنهایت داشبورد بلادرنگ مورد نظر طراحی و پیاده‌سازی شد و از جنبه‌های گوناگون توسط مدل ارزیابی طراحی آزمایش‌ها مورد ارزیابی قرار گرفت. با ارزیابی‌های صورت‌گرفته و با تغییر در پیکره‌بندی آپاچی استورم دریافتیم بلادرنگ‌بودن داشبورد تحت تأثیر قرار نخواهد گرفت. گفتنی است ارزیابی‌های انجام‌گرفته داشبورد را تنها از نظر میزان تأخیر در نمایش داده‌ها بررسی کرده و عواملی چون صحت اطلاعات در نظر گرفته نشده‌اند؛ درنهایت می‌توان اظهار داشت داشبوردِ طراحی و پیاده‌سازی‌شده با استفاده از ابزارهای آپاچی استورم و D3.js پس از ارزیابی از لحاظ بلادرنگ‌بودن و تأیید آن با میانگین زمان تأخیر یک‌دقیقه‌وسی‌ثانیه، نیازمندی‌های بلادرنگ کاربران نهایی آن را می‌تواند تأمین کند.

۷- نتیجه‌گیری

در این پژوهش با استفاده از ابزاری که امروزه برای پیاده سازی راه حل‌های کلان‌داده مطرح می‌شوند، داشبوردی برای تحلیل داده‌های شبکه اجتماعی توییتر طراحی و نمونه‌سازی شد. داشبورد طراحی‌شده با استفاده از ارزیابی‌های انجام‌شده، هدف پژوهش را که بلادرنگ‌بودن داشبورد بود، تأیید کرد؛ اما باید دانست هر پژوهشی که انجام می‌گیرد، کامل نبوده و دارای محدودیت‌هایی است، که در ادامه آنها را بررسی خواهیم کرد. یکی از مهم‌ترین محدودیت‌هایی که این پژوهش با آن روبه‌رو بود، عدم دسترسی به firehose توییتر بود؛ زیرا APIهای فعلی چه از نوع REST و چه Streaming تنها نمونه‌ای از داده‌ها را در اختیار توسعه‌دهندگان قرار می‌دهد. گفتنی است، دسترسی به firehose توییتر مستلزم پرداخت هزینه است؛ همچنین در این پژوهش ابتدا اولویت استفاده از داده‌های کشوری بوده، اما در طی فرآیند پیاده‌سازی دسترسی به چنین داده‌هایی نیز میسر نشد. از دیگر محدودیت‌ها در این پژوهش آن است که از بین ویژگی‌هایی که برای کلان‌داده یاد شد، داده‌هایی که دارای ویژگی‌های سرعت تولید و تنوع باشند، انتخاب شدند و ویژگی حجم داده‌ها در این پژوهش در نظر گرفته نشده است؛ همچنین برای بصری سازی تنها از الگوریتم‌های فیلترینگ و تعداد تناوب استفاده شده و استفاده از الگوریتم‌های داده‌کاوی در حوزه این پژوهش نبوده است. با توجه به یافته‌های پژوهش و مطرح‌شدن

[1] Archived data

[DOI: 10.29252/jsdp.16.4.151]

محدودیت‌ها، پیشنهادهای کاربردی برای طراحی و پیاده سازی موفق داشبورد بلادرنگ عبارتند از: انتخاب دقیق هدف طراحی سامانه، طراحی صحیح معماری سامانه بر مبنای هدف، بررسی داده‌های ورودی مورد نیاز سامانه و اطمینان از حصول دست‌یابی به داده‌ها، توجه به نوع پردازش مورد نیاز (دسته‌ای یا جریانی‌بودن پردازش)، توجه به وجود زیرساخت‌های مناسب برای پیاده‌سازی سامانه؛ درنهایت همان‌طورکه مشخص است، کلان‌داده موضوعی به‌نسبه نو به‌خصوص برای کشور ما بوده است. در میان پژوهش‌گران داخلی به‌دلیل محدودیت‌های زیرساختی و داده‌ای گفته‌شده، پژوهش‌های گسترده و معناداری در این حوزه برای سازمان‌های ایرانی صورت نگرفته است. بنابراین به‌نظر می‌رسد برای پژوهش‌های آینده کلان‌داده می‌تواند موضوعی بسیار نو بوده و به آن از جنبه‌ها و زوایای مختلف پرداخت. در انتها می‌توان برای پژوهش‌های آینده موضوعاتی مانند: پیاده‌سازی داشبورد بلادرنگ بر مبنای داده‌های داخلی (داده‌های سازمان‌های تجاری، تراکنش‌های بانکی و ...)، اجرای الگوریتم‌های پیچیده‌تر مانند الگوریتم‌های یادگیری ماشینی برای استخراج بینش‌های عمیق‌تر، پیاده‌سازی داشبورد بلادرنگ با استفاده از سایر ابزارهای کلان‌داده مطرح در مطالعات و استفاده و تجمیع داده از منابع مختلف، را پیشنهاد کرد.

۸- مراجع 8- References

[1] لوشین، دیوید، *تحلیل‌های کلان‌داده؛ نقشه راه پیاده‌سازی، فناوری و ابزارها*، ترجمه روحانی، سعید؛ حسینی، سمیه، نشر نیاز دانش،۱۳۹۴.

[1] D. Loshin, Big Data Analytics, From Strategic Planning to Enterprise Integration With Tools, Techniques, NoSQL, and Graph, translate by Rouhani, Saeed; Hosseini, Somayeh, 1394. Big data analysis; roadmap for implementation; technology and tools.

[2] T. Kumamoto, H. Wada, T.Suzuki, "Proposal of a system for visualizing temporal changes in impressions from tweets", *International Journal of Pervasive Computing and Communications*, Vol. 11, Iss .2, pp. 193 211, 2015.

[3] X.Wu, X.Zhu, G.Wu, W.Ding, "Data mining with big data." *IEEE transactions on knowledge and data engineering*, vol.26(1), pp. 97-107, 2014.

[4] C.Snijders, U.Matzat, U. D.Reips, " Big Data: big gaps of knowledge in the field of internet science", *International Journal of Internet Science*, vol.7 (1), pp.1-5, 2012.

[5] L.Zhang, A.Stoffel, M.Behrisch, S. Mittelstadt, T. Schreck, R.Pompl, S.Weber, H.Last, D.Keim, "Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems", In Visual Analytics Science and Technology (VAST), IEEE, 2012.

[6] I. Hashem, I.Yaqoob, N. Anuar, S.Mokhtar, A.Gani, K.Khan, "The rise of "big data" on cloud computing: Review and open research issues", Information Systems, 47: 98-115, 2015.

[7] J.Abawajy, "Comprehensive analysis of big data variety landscape", *International journal of parallel, emergent and distributed systems*, vol.30 (1): pp.5-14, 2015.

[8] M. R.Bendre, V. R.Thool, "Analytics, challenges and applications in big data environment: a survey", *Journal of Management Analytics*, vol. 3(3), pp.206-239, 2016.

[9] B.Hu, "A Key-Value Based Application Platform for Enterprise Big Data", in IEEE International Congress on Big Data (BigData Congress), 2014.

[10] B. G.Tudorica, C.Bucur, "A comparison between several NoSQL databases with comments and notes", *In Roedunet International Conference (RoEduNet)*, 2011, 10th, pp. 1-5.

[11] R.Cattell, "Scalable SQL and NoSQL data stores," *Acm Sigmod Record*, 39(4), pp. 12-27, 2011.

[12] R. Hecht, S.Jablonski, "NoSQL evaluation: A use case oriented survey," In Cloud and Service Computing (CSC), *2011 International Conference on 2011*, 2011,pp. 336-341.

[13] N.Leavitt, "Will NoSQL databases live up to their promise?" Computer, 43(2), 2010.

[14] A.Corbellini, C.Mateos, A.Zunino, D.Godoy, S.Schiaffino, "Persisting big-data: The NoSQL landscape," Information Systems, vol.63, pp. 1-23, 2017.

[15] W.Ding, G.Marchionini, "A comparative study of web search service performance," *In Proceedings of the ASIST Annual Meeting*, Vol. 33, pp. 136-140, Learned Information, 1996.

[16] X.Jin, B.W. Wah, X.Cheng, Y.Wang, "Significance and challenges of big data research," *Big Data Research*, vol. 2(2), pp. 59-64, 2015.

[17] R.Agrawal, A.Kadadi, X.Dai, & F.Andres, "Challenges and opportunities with big data visualization." In Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems ,pp. 169-173, ACM, 2015.

[DOI: 10.29252/jsdp.16.4.151]

[18] D.Gruhl, M.Nagarajan, J.Pieper , C.Robson, A.Sheth, "Multimodal social intelligence in a real-time dashboard system," *The VLDB Journal The International Journal on Very Large Data Bases*, vol.19 (6), pp. 825-848, 2010.

[19] B.Flesch, "Design, Development and Evaluation of a Big Data Analytics Dashboard", M.S. thesis, Copenhagen Business School, Denmark, 2014.

[20] O.Hoeber , M.Meseery , K.Kenneth Odoh , Gopi, "Visual Twitter Analytics (Vista) Temporally changing sentiment and the discovery of emergent themes within sport event tweets." Online Information Review, vol.40 (1), pp.25-41, 2016.

[21] A.Weiler, M.Grossniklaus, M. H. Scholl, "Situation monitoring of urban areas using social media data streams," *Information Systems-*, vol.57, pp.129-141, 2016.

[22] B.Yadranjiaghdam, "Developing a Real-time Data Analytics Framework For Twitter Streaming Data.", M.S. thesis, Dept.Computer Science, East Carolina Univ., USA, 2016.

[23] B.Yadranjiaghdam, N.Pool, N.Tabrizi, "A Survey on Real-Time Big Data Analytics: Applications and Tools," In Computational Science and Com-putational Intelligence (CSCI), International Conference on, IEEE, 2016.

[24] P.Córdova, "Analysis of real time stream processing systems considering latency", 2015, Available: www.datascienceassn.org.

بابک سهرابی ریاست دانشکده مدیریت و عضو هیأت علمی گروه مدیریت فناوری اطلاعات دانشکده مدیریت دانشگاه تهران و حوزه‌های پژوهشی ایشان، هوش کسب و کار، تحلیل کلان‌داده و داده‌کاوی است.
نشانی رایانامه ایشان عبارت است از:

bsohrabi@ut.ac.ir

سعید روحانی عضو هیأت علمی گروه مدیریت فناوری اطلاعات دانشکده مدیریت دانشگاه تهران و حوزه‌های پژوهشی ایشان، هوش کسب و کار، تحلیل کلان‌داده و داده‌کاوی است.
نشانی رایانامه ایشان عبارت است از:

SRouhani@ut.ac.ir

طاهره پزشکی تحصیلات خود را در در مقطع کارشناسی و کارشناسی ارشد در رشته فناوری اطلاعات گذرانده است. زمینه پژوهشی مورد علاقه ایشان کلان‌داده و فناوری‌های مرتبط با آن است.
نشانی رایانامه ایشان عبارت است از:

t.pezeshki@ut.ac.ir

[DOI: 10.29252/jsdp.16.4.151]