به نام خدا

مبانی رایانش ابری

تمرین سوم

آشنایی عملیاتی با Hadoop و Mapreduce

طراحی تمرین:

خانم‌ها ستارپور و صبا

استاد درس:

آقای دکتر جوادی

مهلت نهایی ارسال پاسخ:

۶ دی ماه ۱۴۰۰ ساعت ۲۳:۵۹

نکته مهم: دقت کنید که تمدید نخواهیم داشت و صرفا می‌توانید ۱ تا ۱۰ روز از ۲۱ روز مجاز برای تاخیر ارسال نمامی تمرین‌ها در این ترم را استفاده کنید. نتیجه محاسبه بودجه باقیمانده شما اعلام خواهد شد.

بخش اول: نصب و راه اندازی خوشه‌ی Hadoop

در کلاس درس با چارچوب Yarn آشنا شده‌اید. در این تمرین، یک خوشه‌ی Hadoop را با استفاده از سه ماشین مجازی راه اندازی و برنامه‌های Mapreduce را روی آن اجرا می‌کنید.

برای ایجاد ماشین‌های مجازی، نصب Hadoop و راه اندازی خوشه، مراحل ذکر شده در لینک زیر را با دقت دنبال کنید:

https://pnunofrancog.medium.com/how-to-set-up-hadoop-3-2-1-multi-node-cluster-on-ubuntu-20-04-inclusive-terminology-2dc17b1bff19

** در مرحله‌ی ۸ در لینک فوق، فایل tar را از لینک زیر دانلود بفرمایید:

https://mirrors.sonic.net/apache/hadoop/common/hadoop-3.2.2/hadoop-3.2.2.tar.gz

** در مرحله ۱۹ مقدار **replication** را برابر با ۱ قرار دهید.

به نکات زیر توجه داشته باشید:

1 - به ماشین مجازی اول 1 vCPU و 1 GB Ram و 20 GB حافظه دیسک و به ماشین‌های مجازی دوم و سوم 2 vCPU و حافظه‌ی بیشتر (مثلا 2GB) اختصاص دهید.

2 - اگر مراحل را به درستی دنبال کنید، نصب به گونه‌ای انجام می‌شود که ماشین مجازی اول نقش‌های NameNode و RecourceManager و ماشین‌های مجازی دوم و سوم نقش‌های DataNode و NodeManager را به عهده می‌گیرند (با استفاده از دستور jps، صحت این مسئله را بسنجید و از آن اسکرین شات تهیه کنید و در گزارش خود بیاورید).

3 - نیازی نیست از مراحل نصب گزارشی تهیه کنید و در این مرحله کافیست نشان دهید ماشین‌های مجازی، نقش‌های گفته شده را بر عهده گرفته‌اند.

4 - نشان دهید که WebGUI از کامپیوتر شخصی شما قابل دسترسی است.

5 - در WebGUI، از قسمت active nodes چه اطلاعاتی به دست می‌آورید؟ ارتباط این اطلاعات را با منابعی که به ماشین‌های مجازی اختصاص داده‌اید، شرح دهید.

توضیحات dataset:

- این dataset شامل ۱.۷۲ میلیون توییت با مضمون انتخابات امریکا است.
- رکوردهای این dataset دارای ۲۱ ستون هستند.
- اطلاعات موجود درباره‌ی ستون‌های این dataset را می‌توانید در لینک زیر مشاهده کنید:

https://www.kaggle.com/manchunhui/us-election-2020-tweets?select=hashtag_joebiden.csv

- دقت کنید که مجموعه داده‌ای که ما در اختیار شما گذاشته‌ایم با dataset لینک فوق تفاوت دارد و تنها اطلاعات ستون‌ها را می‌توانید از این لینک به دست بیاورید و برای اجرای برنامه لازم است که پوشه‌ی datasets.zip را که همراه با دستور کار برای شما در سایت درس بازگذاری شده است، دانلود کنید و از dataset موجود در آن استفاده کنید.
- توییت‌های موجود در فایل new_hashtag_donaldtrump.csv دارای هشتگ‌های #DonaldTrump و یا #Trump و توییت‌های موجود در فایل new_hashtag_joebiden.csv دارای هشتگ‌های #JoeBiden و یا #Biden هستند. دقت کنید که ممکن است برای مثال توییت‌هایی در فایل new_hashtag_donaldtrump.csv وجود داشته باشند که دارای هشتگ #Biden نیز هستند.
- در برخی از رکوردهای dataset، ممکن است اطلاعات یک ستون وجود نداشته باشد (خالی یا null باشد).

بخش دوم: توسعه و اجرای برنامه‌ی Mapreduce

1 - با استفاده از HDFS CLI، پوشه‌ی user/hadoop/ را در HDFS ایجاد کنید.

2 - فایل datasets.zip را (که همراه با دستورکار در سایت درس آپلود شده است) دانلود و از حالت zip خارج کنید.

3 - دو فایل csv موجود در مسیر /datasets/US_election را با استفاده از HDFS CLI در HDFS مثلا در مسیر /user/hadoop/input با replication 1 بارگذاری کنید. ***دقت کنید که هر دو فایل باید فقط با یک بار اجرای برنامه و به صورت همزمان بررسی شوند.***

4 - یک برنامه mapreduce بنویسید که تعداد کل لایک‌ها و تعداد کل retweet را برای توییت‌های مربوط به هردو کاندیدا، Joe Biden و Donald Trump را حساب کند. به این صورت که در هر خط به ترتیب نام کاندید، تعداد لایک‌ها و در نهایت تعداد retweet چاپ شود.

- **دقت کنید که فایل خروجی شما نباید اطلاعات دیگری را شامل شود.**

5 - یک برنامه mapreduce بنویسید که نشان می‌دهد چه بخشی (درصدی) از توییت‌های مربوط به هر یک از کشورهای زیر به ترتیب درباره هر دو کاندیدا، Joe Biden و Donald Trump هستند و در نهایت تعداد کل توییت‌های مربوط به آن کشور را نیز ذکر کنید.

- لیست کشورهای مورد نظر:

Countries = {America, Iran, Netherlands, Austria, Mexico, Emirates, France, Germany, England, Canada, Spain, Italy}

- دقت کنید که فایل خروجی شما نباید اطلاعات دیگری را شامل شود.
- برای این کار از فیلد country استفاده کنید.
- فیلد country در dataset لزوما شامل مقادیر استاندارد نیست؛ بنابراین برای نوشتن این برنامه لازم است که چک کنید هر یک از نام‌های کشورهای فوق در فیلد country وجود دارند یا خیر. برای مثال اگر مقدار

این فیلد برای یک توییت برابر با "somewhere in iran" بود، این توییت باید لحاظ شود. همچنین این جستجو را به صورت case-insensitive لحاظ کنید.

- فیلدهای فایل خروجی باید به ترتیب برابر با نام کشور (فقط به صورت ذکر شده در لیست داده شده یعنی بدون هیچ کاراکتر اضافی دیگری)، درصد توییت‌هایی که درباره‌ی هر دو کاندیدا بودند، درصد توییت‌هایی که درباره‌ی Joe Biden بودند، درصد توییت‌هایی که درباره‌ی Donald Trump بودند و تعداد کل توییت‌های بررسی شده برای آمارگیری این قسمت، باشند.
  نمونه رکورد خروجی:

```
netherlands	0.33676175170751305	0.26420249096022497	0.39903575733226193	12445
```

- دقت کنید که مقادیر هر یک از فیلدها نیز می‌توانند شامل "," باشند.

6 - یک برنامه mapreduce، با عملکرد و قالب خروجی مشابه برنامه‌ای که در قسمت ۵ نوشتید، بنویسید با این تفاوت که این بار برای تعیین کشوری که توییت از آن ارسال شده است، از طول و عرض جغرافیایی استفاده کنید.

- در این برنامه کافی است تنها توییت‌های مربوط به کشورهای امریکا و فرانسه را مورد بررسی قرار دهید.
- طول و عرض جغرافیایی شهرهای کشورهای آمریکا و فرانسه، به صورت تقریبی به قرار زیر است:
  - آمریکا: ۶۸- > طول جغرافیایی > ۱۶۱.۷۵-
    ۶۴.۸۵ > عرض جغرافیایی > ۱۹.۵
  - فرانسه: ۹.۴۵ > طول جغرافیایی > ۴.۶۵-
    ۵۱ > عرض جغرافیایی > ۴۱.۶
- نتایج حاصل از دو قسمت ۵ و ۶ را با هم مقایسه کنید و علت تفاوت را ذکر کنید.

* توجه داشته باشید تمامی نتایج بدست آمده را همراه با کدهای تمامی قسمت‌ها و گزارش خود باید ارسال کنید.

آنچه که باید ارسال کنید

یک فایل زیپ با نام SID_HW3.zip که شامل موارد زیر است:

- فایل‌های مربوط به کدهای MapReduce و فایل‌های نتایج
- تحلیل نتایج بدست آمده و موارد خواسته شده در تعریف تمرین در قالب یک گزارش مرتب و خوانا

موفق باشید

تیم درس مبانی رایانش ابری