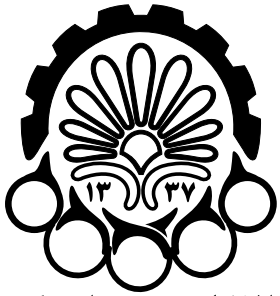
به نام خدا



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

رایانش ابری

تمرین سوم
آشنایی عملیاتی با Hadoop و MapReduce

طراحی تمرین:

خانم ستارپور و آقایان حبیب اله و احمدوند

استاد درس:

آقای دکتر جوادی

مهلت نهایی ارسال پاسخ:

۳۰ اردیبهشت ۱۴۰۱ ساعت ۲۳:۵۹

نکته مهم: دقت کنید که تمدید نخواهیم داشت و صرفاً می‌توانید ۷ روز از ۱۴ روز مجاز برای تاخیر ارسال تمرین‌ها در این ترم را استفاده کنید. به ازای هر روز تاخیر مازاد تا زمان ارائه اسکایپی، ده درصد نمره آن تمرین به عنوان جریمه، از نمره نهایی تمرین کسر می‌شود. ما در اعمال این قاعده جدی هستیم.

گام 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://archive.apache.org/dist/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 ۱ 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-۲۰۲۰-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 دستورکار 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).

گام دوم: توسعه و اجرای برنامه‌ی Mapreduce

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

2- پوشه‌ی datasets.zip را (که همراه با دستورکار در سایت درس آپلود شده است) دانلود و اکسترکت 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 برنامه‌ی MapReduce بنویسید که تعداد پسندیدن (لایک‌ها)، retweet و تعداد source استفاده شده را برای توییت‌های مربوط به Joe Biden ، Donald Trump و هر دو کاندید را حساب کند. به این صورت که در هر خط به ترتیب نام کاندید، تعداد لایک‌ها، تعداد کل retweet و در نهایت تعداد هر source مشخص شده (Web App، iPhone و Android) به ترتیب چاپ شود، مطابق با فرمت زیر:

```
Both Candidate   likes   retweets        Twitter Web App         Twitter for iPhone      Twitter for Android
Donald Trump     likes   retweets        Twitter Web App         Twitter for iPhone      Twitter for Android
Joe Biden        likes   retweets        Twitter Web App         Twitter for iPhone      Twitter for Android
```

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

5- یک برنامه‌ی MapReduce بنویسید که نشان می‌دهد چه بخشی از توییت‌های مربوط به هر یک از ایالت‌های زیر به ترتیب در بازه‌ی بسته ۹ صبح تا ۵ عصر درباره هر دو کاندیدا، Joe Biden و Donald Trump هستند و در نهایت تعداد کل توییت‌های مربوط به آن ایالت در بازه زمانی مشخص شده را نیز ذکر کنید.

- لیست ایالت‌های مورد نظر:

States = {New York, Texas, California, Florida}

- دقت کنید که فایل خروجی شما نباید اطلاعات دیگری را شامل شود.
- برای ایالت‌ها از فیلد state و زمان توییت از فیلد created_at استفاده کنید.
- این جستجو را به صورت **case-insensitive** انجام دهید.
- دقت کنید که مقادیر هر یک از فیلدها نیز می‌توانند شامل ”,“ باشند.
- فیلدهای فایل خروجی باید به ترتیب برابر با نام ایالت (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 در بازه زمانی مشخص شده برای آمارگیری این قسمت، باشند.

نمونه رکورد خروجی:

```
new york        0.26102359237205097     0.36330745458656816     0.37566895304138087     13267
```

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 ایالتی که توییت از آن ارسال شده است، از طول و عرض جغرافیایی استفاده کنید.

- در این برنامه کافی است تنها توییت‌های مربوط به ایالت نیویورک و کالیفورنیا را مورد بررسی قرار دهید.
- طول و عرض جغرافیایی این دو ایالت، به صورت تقریبی به قرار زیر است:
  - نیویورک: ۷۱/۷۵۱۷- > طول جغرافیایی > ۷۹/۷۶۲۴-

    ۴۵/۰۱۵۳ > عرض جغرافیایی > ۴۰/۴۷۷۲
  - کالیفورنیا: ۱۱۴/۱۳۱۵- > طول جغرافیایی > ۱۲۴/۶۵۰۹-

    ۴۲/۰۱۲۶ > عرض جغرافیایی > ۳۲/۵۱۲۱
- **نتایج حاصل از دو قسمت ۵ و ۶ را با هم مقایسه کنید و علت تفاوت را ذکر 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