به نام خدا

درس رایانش ابری

تمرین سوم

آشنایی عملیاتی با **Hadoop**

طراحی تمرین:

آقایان ذوالفقاری و حمدی

استاد درس:

آقای دکتر جوادی

مقدمه

هدف این تمرین، اجرای **MapReduce** بر روی خوشه **Hadoop** است. از آنجایی که در تمرین قبل کار با داکر را یاد گرفته‌اید و از طرفی راه‌اندازی خوشه **Hadoop** به صورت دستی مراحل زیادی دارد، در این تمرین قرار هست **Hadoop** را با استفاده از داکر راه‌اندازی کنید و سپس بر روی آن **MapReduce** های مختلفی اجرا کنید.

گام اول: راه اندازی خوشه Hadoop

برای این تمرین، یک فایل داکر کامپوز آماده شده است که برای شما پیش‌نیازهای زیرساختی تمرین را فراهم می‌کند که شامل کانتینرهای زیر است:

- **NameNode**
- **DataNode**
- **NodeManager**
- **ResourceManager**
- **HistoryManager**

برای راه‌اندازی **Hadoop** و بالا آوردن کانتینرهای ذکر شده، کافی است دستور **docker-compose up -d** را اجرا کنید.

موارد زیر را در فایل گزارش نمایش دهید:

- **نمایش کانتینرهای ایجاد شده**
- **توضیح وظیفه هرکدام از کانتینرها در Hadoop**
- **با استفاده از دستور jps در هر کانتینر، صحت نقش آن کانتینر در Hadoop را بررسی کنید و اسکرین شات آن را بیاورید.**

- **نمایش WebUI برای NameNode (localhost:9870) و فایل سیستم آن**

گام دوم: توضیحات دیتاست

- این دیتاست شامل 200000 توییت با مضمون انتخابات امریکا است.
- رکوردهای این دیتاست دارای 21 ستون هستند.
- در برخی از رکوردهای دیتاست، ممکن است اطلاعات یک ستون وجود نداشته باشد (خالی یا مقدار **Null** باشد)

برای استفاده از دیتاست کافی است فایل **dataset.csv** را در کانتینر **NameNode** قرار دهید.

گام سوم: توسعه و اجرای برنامه‌ی MapReduce

1. با استفاده از **HDFS CLI**، پوشه‌ی **user/root/input** را در **HDFS** ایجاد کنید.
2. فایل **dataset.csv** را با استفاده از **HDFS CLI** در **HDFS** در پوشه‌ی **input** قرار دهید.

حال قرار است سه برنامه **MapReduce** بنویسید که از دیتاستی که در مسیر **HDFS** قرار داده شده است، استفاده کند.

3. یک برنامه **MapReduce** بنویسید که تعداد **likes**، **retweet** و **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.

4. یک برنامه MapReduce بنویسید که نشان می‌دهد چه بخشی از توییت‌های مربوط به هر یک از ایالت‌های زیر به ترتیب در بازه‌ی بسته نه صبح تا پنج عصر درباره هر دو کاندید، 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}

نکات:

- 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 کاندید 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 خروجی:

```
new york        0.26102359237205097     0.363307454586568l6     0.37566895304138087     13267
```

5. یک برنامه MapReduce، با عملکرد و قالب خروجی مشابه برنامه‌ای که در قسمت 4 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.

نکات:

- در این برنامه کافیست تنها توییت‌های مربوط به ایالت نیویورک و کالیفورنیا را مورد بررسی قرار دهید.
- طول و عرض جغرافیایی این دو ایالت، به صورت تقریبی به صورت زیر است:
  - **نیویورک:**

    -71.7517 > طول جغرافیایی > -79.7624

    45.0153 > عرض جغرافیایی > 40.4772
  - **کالیفورنیا:**

    -114.1315 > طول جغرافیایی > -124.6509

    42.0126 > عرض جغرافیایی > 32.5121
- نتایج حاصل از دو قسمت 4 و 5 را با هم مقایسه کنید و در صورت تفاوت، علت آن را توضیح دهید.

نکات اجرای MapReduce:

- برای نوشتن کدهای **MapReduce** می‌توانید از زبان جاوا و یا پایتون استفاده کنید. (نسخه پایتون موجود در همه‌ی کانتینرها 3.5.3 است)
- برای اجرای کدهای **MapReduce** در **Hadoop**، کافی است در کانتینر **NameNode** آن را با استفاده از دستور **Hadoop** اجرا کنید.

موارد زیر را در فایل گزارش نمایش دهید:

- **خروجی هر یک از برنامه‌های MapReduce**
- **نمایش WebUI برای NameNode (localhost:9870) و فایل سیستم آن (دیتاست و خروجی برنامه‌های MapReduce)**

نکات مربوط به تحویل تمرین

- تمرین شما تحویل اسکایپی خواهد داشت. بنابراین از استفاده از کدهای یکدیگر یا کدهای موجود در وب که قادر به توضیح عملکرد آنها نیستید، بپرهیزید!
- ابهامات خود را در سایت درس یا در گفت‌وگوی خصوصی تلگرام با تدریس‌یاران مطرح کنید و ما در سریع‌ترین زمان ممکن به آنها پاسخ خواهیم داد.

آنچه که باید ارسال کنید

یک فایل زیپ با نام sid_HW3.zip که شامل موارد زیر است:

1. کدهای MapReduce و نتایج آن
2. گزارش که حداقل باید شامل موارد مطرح شده در توضیحات تمرین (به همراه اسکرین‌شات) باشد

موفق باشید

تیم درس مبانی رایانش ابری