دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس کلان داده

نیمسال دوم سال تحصیلی ۹۹-۹۸

تمرین شماره ۱

Hadoop

اسفندماه ۱۳۹۸

مقدمه

هدف از این تمرین آشنایی شما با Hadoop و چارچوب پردازشی MapReduce در قالب انجام چند تمرین بر اساس محتویات فصل سوم کتاب (Big Data Analytics)است . در صورت تمایل می‌توانید فایل ویدئوی آپلود شده در این خصوص که توضیح اسلایدهای مربوط به این بخش است را مشاهده نمایید.

با توجه به مسایل و مشکلاتی که هنگام نصب و اجرای هدوپ معمولاً اتفاق میفتد، تالار گفتگویی ذیل همین تمرین برای اشتراک مسایل و راهنمایی در نظر گرفته شده است که می‌توانید در این تالار، سوالات و مسایل خود را مطرح کرده و یا به سایرین در حل مشکلاتشان کمک کنید.

بخش اول) اجرای هدوپ

در بخش اول، باید بتوانید هدوپ را در یک سرور(ترجیحاً) لینوکس بالا آورده و فایل‌های AccessLog را که در بخش منابع این تمرین قرار گرفته است، به بخش مدیریت فایل آن یعنی HDFS منتقل کنید.

برای کار با هدوپ، می‌توانید :

- مستقیماً آنرا بر روی لینوکس[1] (و حتی ویندوز[2]) نصب کنید.
- از ایمیج‌های آماده مانند CDH استفاده کنید (طبق راهنمای کتاب)
- از داکر[3] برای اجرای آن استفاده کنید.

انتخاب روش بر عهده خودتان است اما طبق تجربه، اگر بتوانید با داکر کار کنید و با دانلود حجم بالای ایمیج‌های آن (حدود پنج تا 10 گیگابایت) مشکلی نداشته باشید، بی‌دردسرترین روش را انتخاب کرده‌اید.

نکته اول : اگر مستندات کلاس برگزار شده در خصوص داکر را داشته باشید یا در آن شرکت کرده باشید، به راحتی کار با داکر را می‌توانید شروع کنید. برای کار با داکر در ویندوز می‌توانید نرم افزار Docker Desktop را نصب کنید (نوع کانتینرها را لینوکسی انتخاب کنید) تا دستورات آن در خط فرمان ویندوز قابل اجرا باشد و یا یک لینوکس سرور به صورت مجازی نصب کنید و داکر را بر روی آن اجرا کنید.

نکته دوم : در صورت استفاده از CDH، نسخه 5 آن هم برای این تمرین مناسب است.

نکته سوم : برای انتقال فایل‌ها از سیستم خودتان به سرور هدوپ، می‌توانید از نرم افزار Bitvise SSH Client استفاده کنید. بعد از انتقال فایل‌ها به سرور هدوپ و با اتصال به خط فرمان سرور هدوپ، از دستور hadoop fs – copyfromlocal (یا دستور put در Hadoop fs) برای انتقال فایل اکسس‌لاگ به HDFS و احیاناً ایجاد پوشه در آن (ترجیحاً با نام data) استفاده کنید.

توضیح اینکه فایل سیستم سروری که هدوپ در آن در حال اجراست با فایل سیستم داخلی خود هدوپ یعنی HDFS متفاوت است.

موارد مورد نیاز ارسالی در این بخش :

فرآیند نصب و اجرای هدوپ و نحوه انتقال فایلها و اطمینان از انتقال آنها را در فایل ورد گزارش (که البته درانتهای کار، PDF آنرا ارسال خواهید کرد)، مستند کنید.

[1] https://acadgild.com/blog/hadoop-3-x-installation-guide
[2] https://dev.to/awwsmm/installing-and-running-hadoop-and-spark-on-windows-33kc
[3] https://hub.docker.com/r/cloudera/quickstart

بخش دوم) دست گرمی با WordCount

در این بخش از تمرین، هدف، شمردن تعداد تکرار کلمات در یک فایل متنی به عنوان یک مثال کلاسیک در حوزه پردازش کلان‌داده به کمک کتابخانه MRJob در پایتون (فصل سوم کتاب) است.

برای اینکار ابتدا فایل متن کتاب جنگ و صلح تولستوی را که از این آدرس[1]، دانلود نموده و سپس در HDFS بارگذاری نمائید، در ادامه با استفاده از Hadoop و مبتنی بر مدل برنامه‌نویسی نگاشت/کاهش دو فایل Mapper و Reducer ایجاد نمائید و با استفاده از آنها تعداد تکرار کلمات در فایل متنی داده شده را بشمارید و ده تای پرتکرار را نمایش دهید. نمونه ای از آنچه مدنظر است (قبل از مرحله مرتب‌سازی) را در زیر می توانید مشاهده کنید.

```
[('we', 919),
 ('for', 948),
 ('is', 1507),
 ('symbol', 1540),
 ('and', 1575),
 ('to', 1737),
 ('in', 1814),
 ('a', 1949),
 ('of', 2993),
 ('the', 5039)]
```

پیش پردازش‌های لازم شامل یکسان سازی حروف کوچک و بزرگ و حذف علامت های چسبیده به حروف مانند ویرگول را در نظر داشته باشید (به طور مثال The و the یک کلمه مشابه هستند و لازم است در شمارش یک کلمه محاسبه شوند و یا ,The و The نیز تفاوتی با هم ندارند.)

برنامه را در سه حالت زیر اجرا کنید :

- در حالت لوکال و بدون استفاده از هدوپ (معمولاً برای تست اولیه برنامه‌ها استفاده می‌شود)
- با استفاده از هدوپ و با آدرس دهی فایلهای ورودی از سرور اصلی
- با استفاده از هدوپ و با آدرس دهی فایلها از HDFS

برای هر یک حالات فوق مثالی از نحوه فراخوانی کدها در زیر ارائه شده است :

```
python mr_word_count.py  /data/war_and_peace_tolstoy.txt
python mr_word_count.py  -r hadoop /data/war_and_peace_tolstoy.txt
python mr_word_count.py -r hadoop hdfs:///data/war_and_peace_tolstoy.txt
```

در مثال دوم فایل ورودی از سیستم محلی خوانده شده و برای اجرا به یک آدرس موقت در HDFS منتقل می‌شود.
در مثال آخر، فایل ورودی از پوشه data در hdfs خوانده شده است.

[1] http://www.gutenberg.org/files/2600/2600-0.txt

موارد مورد نیاز ارسالی در این بخش :

در این بخش، فایل کدها را در پوشه‌ای با نام Sec2 و نتیجه بدست آمده از اجرای کد در هر یک از سه حالت و تفاوت‌های مشاهده شده را در فایل گزارش، وارد کنید.

بخش سوم) پردازش لاگ‌ها - یافتن محبوب‌ترین عکس‌ها

فایل‌های access_log که در بخش اول تمرین به HDFS منتقل شده‌اند شامل اطلاعات دسترسی بازدیدکنندگان یک سایت هستند. این فایل‌ها شامل یک سری رکورد هستند که هر رکورد به ترتیب حاوی IP کاربر، تاریخ بازدید یک یوآرال یا درخواست آن از سرور ، یوآرال صفحه یا منبع درخواست شده، کد HTTP نتیجه و نهایتاً یک عدد است که در زیر سه خط از این فایل برای نمونه آورده شده است :

10.211.47.159 - - [03/Jan/2010:18:19:24 -0800] "GET /images/frontpagepics/0000/0012/Thirst700.jpg HTTP/1.1" 304 -

10.48.89.142 - - [03/Jan/2010:18:22:30 -0800] "GET /robots.txt HTTP/1.1" 404 208

10.48.89.142 - - [03/Jan/2010:18:22:47 -0800] "GET /show_film.php3?id=2991 HTTP/1.1" 404 186

با ایده گرفتن از سه مثال Top-N ، Filtering و Binning که درفصل ۳ کتاب درسی، با کتابخانه MRJob پایتون پیاده سازی شده است و با استفاده از فایل‌های اکسس لاگ بخش اول، خواسته زیر را با هدوپ دست آورید :

۱۰ پردرخواست‌ترین عکس‌های درخواست شده (Top 10) به ازای هر سال را بیابید.

توضیح اینکه تنها به دنبال یوآرال هایی هستیم که در آنها png. یا jpg. به کار رفته و با images/newspics شروع می‌شوند. به ازای این سطرهای خاص در فایل AccessLog ، نام عکس باید استخراج شده و پرتکرارترین عکس‌ها در هر سال محاسبه شود.

موارد مورد نیاز ارسالی در این بخش :

کدهای نوشته شده را در پوشه‌ای با نام Sec3 قرار دهید و توضیحات لازم در مورد کدها و خروجی تولید شده را در فایل گزارش وارد کنید.

بخش چهارم) پردازش لاگ‌ها - استخراج آمار ماهیانه

در این بخش میخواهیم، آمار ماهیانه بازدید را برای هر سه سال موجود در فایل، محاسبه و نمودار مقایسه‌ای آنها را رسم کنیم. منظور از هر بازدید، هر آی‌پی به ازای هر روز است. یعنی اگر فردی روزی یک‌بار یا بیشتر به سایت ما مراجعه کرده و صفحات و منابع مختلفی را از سرور درخواست کرده باشد، آنرا تنها یک بازدید به حساب می‌آوریم.

بعد از محاسبه آمار ماهیانه بازدید هر سه سال، نمودار ستونی مربوطه را می توانید به کمک اکسل رسم کنید.

اگر بخواهیم بدانیم کدام کاربر(IP)، بیشترین زمان را در سایت ما سپری کرده است (Top 10 for each year)، چه کاری باید انجام دهیم ؟ دقت کنید که زمان بازدید برای هر سشن را باید محاسبه کنید و نهایتاً به ازای هر کاربر، این زمان‌ها را با هم جمع بزنید.

موارد مورد نیاز ارسالی در این بخش :

فایل کدها را در پوشه‌ای با نام Sec4 قرار داده، توضیحات و نمودار به دست آمده را در فایل گزارش وارد کنید.

بخش پنجم) نوشتن Partitioner سفارشی - امتیاز اضافی

در این بخش مجددا تمرین بخش چهارم را انجام دهید ولی با این تفاوت که این بار تعداد Reducer های شما باید دقیقا برابر 12 باشد که هر Reducer وظیفه پردازش یک ماه مشخص را برعهده دارد. در این مرحله باید تابع Partitioner را بازنویسی کنید. شاید راحت‌تر باشد که با جاوا و طبق این آموزش، این کار را انجام دهید :

https://www.tutorialspoint.com/map_reduce/map_reduce_partitioner.htm

موارد مورد نیاز ارسالی در این بخش :

در این بخش فایل کدها و نتیجه بدست آمده از اجرای کد را مشابه فوق ارسال نمایید. تغییرات لازم برای تعیین تعداد Reducer ها و الگوریتم مورد استفاده را توضیح دهید.

نکات پیاده سازی

- در این تمرین فقط مجاز به استفاده از زبان برنامه نویسی Python خواهید بود.(غیر از بخش پنجم که مجازید از جاوا یا هر زبان دیگری استفاده کنید)
- پیش پردازش های لازم را فراموش نکنید!
- استفاده یا عدم استفاده از Docker اختیاری است ولی توصیه می گردد برای راحتی انجام کار از Docker استفاده نمائید.
- تمامی کدهای ارسالی تحت پلتفرم Hadoop اجرا خواهند شد ، از صحت کدها اطمینان حاصل نمائید ، در غیر این صورت نمره بخش مربوطه را از دست خواهید داد.

نکات تحویل

- مهلت ارسال این تمرین تا **۲۹ اسفند** خواهد بود.
- انجام این تمرین به صورت یک نفره می باشد.
- می توانید تمرین را حداکثر با یک هفته تاخیر ارسال نمائید ، نحوه محاسبه تاخیر نیز به این شکل خواهد بود که به ازای هر روز تاخیر ۱۵ درصد از نمره تمرین کسر خواهد شد.
- بعد از پایان مهلت ارسال تمرین، تمرین تحویل حضوری نیز خواهد داشت ، که زمان آن متعاقبا از طریق سامانه مدیریت دروس اعلام خواهد شد.
- لطفا در روز تحویل حضوری کدهای خود را آماده اجرا داشته باشید ،دقت نمائید که حق تغییر کدهای ارسالی را نخواهید داشت و همچنین افرادی که تمرین خود را تا قبل از تاریخ اعلام شده در سامانه آپلود نکرده باشند حق تحویل حضوری نخواهند داشت.
- می توانید برای پاسخ تمرین ها در اینترنت جستجو کنید اما وجود تشابه غیرمنطقی بین گزارش ها و کدهای ارسالی تقلب محسوب شده و نمره تمرین تمامی افراد شرکت کننده در آن صفر در نظر گرفته خواهد شد.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه ای برخوردار است، لطفا تمامی مواردی که در شرح تمرین از شما خواسته شده را در گزارش ذکر نمائید.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW1_[Lastname]_[StudentNumber].zip

در صورت وجود ابهام یا سوال می توانید از طریق رایانامه های زیر با دستیاران آموزشی تماس بگیرید.

smbanaei@ut.ac.ir alikarimi120@gmail.com