به نام خدا

مبانی رایانش ابری (نیم سال دوم تحصیلی ۹۹-۰۰)

تمرین شماره ۲: نصب و راه‌اندازی Apache Hadoop Yarn، نوشتن و اجرای برنامه‌های MapReduce

آخرین تاریخ آپلود پاسخ در courses: ۷ خرداد ۱۴۰۰، ساعت ۲۳:۵۹

در جلسات کلاس درس با چهارچوب Hadoop Yarn به شکل کامل اشنا شدید و نحوه نصب و اجرای برنامه‌های MapReduce توضیح داده شد. در این تمرین یک Hadoop Cluster با استفاده از سه ماشین مجازی راه‌اندازی کرده و برنامه‌های MapReduce که در ادامه ذکر می شوند را بر روی آن اجرا می‌کنید. مراحل زیر را گام به گام انجام دهید و نتایج را خلاصه‌وار در گزارش خود بیاورید. این تمرین، تحویل **اسکایپی** خواهد داشت. تسلط و ارائه شما در تحویل اسکایپی بسیار حائز اهمیت است.

1 - ایجاد سه ماشین مجازی Ubuntu(20.4 – 18.04) با استفاده از VirtualBox. دقت کنید که به VM1، 1 vCPU و 1 GB Ram و 20 GB حافظه دیسک بدهید اما به VM2 و VM3، 2 vCPU و حدالمقدور حافظه بیشتر (2 GB). همانطور که در کلاس بیان شد، در ابتدا می‌توانید تمامی مراحل نصب Hadoop را بر روی VM1 ایجاد کنید و سپس با استفاده از clone، ماشین‌های مجازی بیشتری را به خوشه Hadoop اضافه کنید.

2 - نصب Apache Hadoop Yarn را به گونه‌ای انجام دهید که VM1، نقش‌های NameNode و ResourceManager را بر عهده بگیرد و VM2 وVM3 نقش های DataNode و NodeManager را بر عهده بگیرند. (می‌توانید از تمام منابع موجود در اینترنت استفاده کنید. یکی از لینک های مفید نیز برای راحتی کار معرفی شده است):

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

3 - همانطور که در اسلایدهای آموزش نصب Hadoop بیان شده است، با اضافه کردن ScreenShot به گزارش خود نشان دهید که WebGUI برای HDFS و Hadoop از کامپیوتر شخصی شما (host) قابل دسترس است. توضیح دهید که HDFS GUI چه اطلاعاتی را نشان می دهد. چه اطلاعاتی را از فضای دیسک قابل دسترس نشان می دهد و همچنین Hadoop Web GUI چه اطلاعاتی از active nodes نشان می دهد و توضیح دهید رابطه این اطلاعات با منابعی که به ماشین های مجازی تخصیص داده‌اید، چیست. **نیازی به طولانی بودن توضیحات نیست (می‌توانید توضیحات را در قالب جدول بیان کنید).**

4 - در این گام با HDFS CLI آشنا شده سپس پوشه /user/Hadoop را در HDFS ایجاد کنید (اجرای دستورات در VM1). سپس فایل test.txt را در VM1 ایجاد کرده (محتویات دلخواه از کلمات) و سپس فایل را با استفاده از HDFS CLI به

HDFS بارگذاری کنید (upload). سپس با استفاده از HDFS CLI نشان دهید که این فایل با موفقیت بارگذاری شده است و در web GUI نیز قابل دسترسی است.

5 - برنامه WordCount را با استفاده از زبان Java و راهنمایی گام به گام نشان داده شده در لینک زیر بر روی فایل مثالی مرحله ۴ اجرا کنید و با اضافه کردن ScreenShot به گزارش خود نشان دهید که برنامه نتیجه درست را اجرا کرده است. متن این برنامه را نیز در فایل‌های ارسالی خود قرار دهید.

https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

6 - ابتدا DataSet موجود در لینک زیر را دانلود کنید و در HDFS بارگذاری کنید. شما بایستی هر چهار زیر دانلود و در HDFS بارگذاری کنید (replication factor را برابر با یک قرار دهید). سپس تعداد جرم‌هایی که در هر district (به معنی ناحیه) رخ داده است را با استفاده از یک برنامه MapReduce محاسبه کنید. فایل خروجی بایستی دارای فرمت زیر باشد و **ناحیه‌ها بایستی از تعداد جرم بیشتر به کمتر مرتب شده باشند**:

File name: crime_count_per_district.csv

File structure:

// District, crime count (just to show the structure; your file should be numeric only)

…, …

فایل پاسخ برنامه MapReduce نیز را ارسال کنید. دقت کنید که برنامه همه شما دانشجویان عزیز بایستی خروجی که صحیح و را تولید کند. تیم درس این خروجی صحیح را برای انجام مرحله ارزیابی و نمره‌دهی تولید خواهد کرد.

لینک دانلود Dataset (این لینک را با فیلتر شکن باز کنید):

https://www.kaggle.com/currie32/crimes-in-chicago

فایل‌هایی که باید دانلود کنید (حجم فایل‌های زیپی که دانلود می‌کنید به مراتب کمتر است.)

Chicago_Crimes_2001_to_2004.csv(453.32 MB)

Chicago_Crimes_2005_to_2007.csv(449.45 MB)

Chicago_Crimes_2008_to_2011.csv(646.3 MB)

Chicago_Crimes_2012_to_2017.csv(349.81 MB)

دقت کنید که بایستی پیکربندی job ارسال شده را به گونه‌ای تغییر دهید که **چهار** map tasks تعریف شود (رخداد این موضوع را در گزارش و در موقع ارائه اسکایپی خود نشان دهید). محدودیتی روی تعداد reduce task وجود ندارد.

District - Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.

نحوه تحویل تمرین دوم

1- موارد زیر را در قالب یک فایل زیپ با نام «HW2_student_id.zip» در صفحه درس بارگذاری کنید.

- گزارش شما که بایستی از کیفیت مناسبی برخوردار باشد و از تکرار یا بی نظمی پرهیز کند.
- کدهای متن برنامه‌های MapReduce که می‌نویسید و همچنین خروجی برنامه MapReduce اصلی (شمارش تعداد جرم‌ها) باید پیوست شده باشد. از شما دانشجویان عزیز انتظار رعایت همه اصول و انتخاب بهترین رویکردها را داریم. کدهای شما نباید copy-paste از راه‌حل‌های موجود در وب یا دیگر دانشجویان باشد و بایستی بتوانید راه حل خودتان را با تسلط کامل به دستیاران آموزشی توضیح دهید.

2- دستیاران آموزشی علاوه بر بررسی گزارش‌ها و کدهای برنامه، از طریق اسکایپ، تمرینات را به صورت اجرای زنده از شما تحویل خواهند گرفت. بنابراین بسیار مهم است به انجام کارهای خواسته شده در هنگام ارائه اسکایپی، تسلط داشته باشید.

جریمه دیرکرد

هرروز تاخیر در ارسال تمرین ۱۰٪ نمره منفی خواهد داشت. امکان اپلود تمرین تنها تا ۵ روز از تاریخ تعیین شده ممکن خواهد بود.

جریمه تقلب

۱- همه بایستی که خود تمرین را انجام دهند و هرگونه تقلب یا ارسال کار دیگران یا کارهای موجود در وب غیرقابل پذیرش بوده و عواقب شدیدی خواهد داشت. دانشجویان بی شک می توانند از راهنمایی های موجود در وب یا کتابخانه های کمکی استفاده کنند ولی باید همه منابع و کتابخانه ها به صراحت ذکر شده باشند.

۲- گروه حل تمرین تمام تلاش خود را برای شناسایی تقلب های احتمالی خواهند کرد تا در نهایت یک ارزیابی عادلانه از همه دانشجویان عزیز داشته باشیم.

در نهایت، هرگونه سوال در مورد تمرین و بخش های آنها را تنها از طریق سایت درس و ایجاد مباحثه با عناوین مرتبط مطرح بفرمایید.

تندرست و موفق باشید.

تیم تدریسیاری مبانی رایانش ابری