به نام خدا

مبانی رایانش ابری (نیمسال اول تحصیلی ۹۹-۴۰۰)

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

آخرین تاریخ اپلود پاسخ در courses:

ساعت ۲۳:۵۹ ۵ دی ماه ۱۳۹۹

بخش۱

در کلاس درس بحث شد که Hadoop Map Task خروجی خود را در local disk می‌نویسد. بنابراین برای اماده کرده ورودی فاز Reduce داده‌ها بایستی از local disk خدمتگزارها[1] خوانده شوند تا مرحله Shuffling انجام شود. توضیح دهید که این نحوه طراحی Apache Hadoop چه سرباری را از منظر زمان اجرای یک برنامه MapReduce ایجاد می‌کند. حداقل در دو پارگراف (هر کدام چندین جمله) توضیح دهید که Apache Spark چیست و چگونه مشکل کارائی مطرح شده در قسمت ابتدائی این سوال را حل کرده است؟ علاوه بر این یک مقایسه کامل‌تر از تفاوت اجرای یک برنامه MapReduce در Hadoop و Spark ارائه دهید.

بخش۲

در جلسات کلاس درس با چهارچوب Hadoop Yarn به شکل کامل آشنا 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 را بر روی آن اجرا می‌کنید. مراحل زیر را گام به گام انجام دهید و نتایج را در گزارش خود بیاورید.

۲-۱ ایجاد سه ماشین مجازی Ubuntu 20.4 با استفاده از VirtualBox. دقت کنید که به VM1 ، 1 vCPU و 1GB Ram و 20GB حافظه دیسک تخصیص بدید. اما به VM2 و VM3، 2 vCPUs و حدالمقدور حافظه بیشتر (2GB Ram). همانطور که در کلاس بیان شد، در ابتدا می توانید تمامی مراحل نصب Hadoop را بر روی VM1 ایجاد کنید و سپس با استفاده از عملیات Clone دو ماشین مجازی جدید از VM1 ایجاد کنید و منابع بیشتری به آنها تخصیص بدهید.

۲-۲ نصب Apache Hadoop Yarn را به گونه‌ای انجام دهید که VM1، نقش‌های NameNode و ResourceManager را بر عهده بگیرد و VM2 و VM3، نقش‌های DataNode و NodeManager را بر عهده بگیرند.

[1] servers

۲-۳ همانطوری که در اسلایدهای اموزش نصب Hadoop بیان شده است، با اضافه کردن ScreenShot به گزارش خود نشان دهید که Web GUI برای HDFS و Hadoop از کامپیوتر شخصی شما (host) قابل دسترسی است. توضیح دهید که HDFS GUI چه اطلاعات کلی از فضای دیسک قابل دسترس را نشان می‌دهد و همچنین Hadoop Web GUI چه اطلاعات کلی از active nodes نشان می‌دهد و توضیح دهید رابطه این اطلاعات با منابعی که به ماشین‌های مجازی تخصیص داده‌اید، چیست.

۲-۴ در این گام، با HDFS CLI اشنا شده سپس پوشه /user/hadoop را در HDFS ایجاد کنید (اجرای دستورات در VM1). سپس فایل test.txt را در VM1 ایجاد کرده (با محتویات دلخواه از کلمات) و سپس فایل را با استفاده از HDFS CLI به HDFS بارگذاری کنید. سپس با استفاده از HDFS CLI نشان دهید که این فایل با موفقیت بارگذاری شده است. همچنین از طریق اضافه کردن ScreenShot به گزارش خود نشان دهید که پوشه ایجاد شده و فایل بارگذاری شده در HDFS WEB GUI نیز قابل دسترسی است.

۲-۵ برنامه WordCount را با استفاده از زبان Java و راهنمایی گام به گام داده شده در لینک زیر بر روی فایل مثالی مرحله ۲-۴، اجرا کنید و با اضافه کردن ScreenShot به گزارش خود نشان دهید که برنامه نتیجه درست را ایجاد 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

۲-۶ تا به اینجا مواردی را انجام داده‌اید که در کلاس درس بحث شد. بخش اصلی پیاده‌سازی ضرب دو ماتریس با استفاده از MapReduce است که بر روی Hadoop Cluster ایجاد شده در مراحل قبلی، اجرا می شود. ورودی برنامه MapReduce شما بایستی مسیر فایل دو ماتریس A و B باشد. A و B بایستی که فایل‌های متنی باشند و بر روی HDFS بارگذاری شوند. برای تست درستی برنامه خود، ابتدا به شکل دقیق نشان دهید که برنامه MapReduce به درستی دو ماتریس 2*2 را در همدیگر ضرب می‌کند. بدین منظور دو ماتریس متنی مثالی 2*2 در VM1 ایجاد کنید و انها را به HDFS بارگذاری کرده و برنامه MapReduce را اجرا کنید. بایستی نشان دهید که خروجی به درستی دو ماتریس ورودی را در هم ضرب کرده است. در مرحله نهایی، دو ماتریس تصافی 1000*1000 ایجاد کنید و مجددا انها را به HDFS بارگذاری کرده و برنامه خود را اجرا کنید. با استفاده از ابزارهای موجود، مقدار استفاده شده از CPU و حافظه کامپیوتر خود را در خلال اجرای این برنامه گزارش کنید. همچنین زمان اجرای برنامه خود را گزارش کنید.

نحوه تحویل تمرین ۲

۱. یکی از اعضای گروه، موارد زیر را در قالب یک فایل زیپ با نام «group_id_student_id1_student_id2_hw2» در صفحه درس اپلود کند. شماره گروه را از فایل اکسلی که برای تشکیل گروه‌ها استفاده شد، بازیابی کنید.

- گزارش که بایستی شامل پاسخ به بخش اول و گزارش انجام گام‌های مختلف بخش دوم باشد. گزارش شما بایستی که از کیفیت خوب برخوردار بوده و از تکرار یا بی نظمی پرهیز کنید. ***اولین بخش در گزارش جدولی است که تعیین***

می‌کند هر عضو گروه چه کارهایی را انجام داده است. این تقسیم کار بایستی در زمان تحویل حضوری قابل راستی آزمایی باشد.

- کدهای متن برنامه‌های MapReduce که نوشته‌اید. انتظار حرفه‌ای بودن از شما دانشجویان عزیز و رعایت همه اصول و بهترین رویکردها (best practice) در برنامه‌نویسی را داریم. کدهای شما نمی‌تواند copy-paste از راه‌حل‌های موجود در وب باشد و تنها می توانید از آنها ایده بگیرید و بایستی که بتوانید راه حل را خود را با تسلط کامل به دستیاران آموزشی توضیح دهید.

۲. دستیاران اموزشی 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 که هر دو عضو گروه به پروژه تسلط داشته باشند و انجام تمرین یا حداقل اجرای کد در سیستم هر دو عضو گروه توصیه می‌شود (اگر چنین امکانی وجود داشته باشد). تسلط هر دو عضو گروه در ارائه نقش مهمی در نمره نهایی شما خواهد داشت. انتظار می‌رود عضوی از یک گروه که تسلط بیشتری به این تمرین دارد، با همکاری موثر وکمک به عضو دیگر، نقش مهمی در آموزش جمعی حاصل از این کلاس، ایفا کند.

جریمه دیرکرد

هر روز 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 گروه‌ها بایستی که خود تمرین را انجام دهند و هرگونه تقلب یا ارسال کار دیگران یا کارهای موجود در وب که بخش برنامه‌نویسی تمرین را به شکل جزئی یا کلی انجام داده است، غیرقابل پذیرش و عواقب شدیدی خواهد داشت. دانشجویان بی شک می‌توانند از راهنمای موجود در وب یا کتابخانه‌های کمکی استفاده بکنند تا جایی که همه منابع و کتابخانه‌ها کمکی به صراحت ذکر شده باشد.

۲. بنده و گروه حل تمرین تمام تلاش خود را برای شناسایی تقلب‌های احتمالی خواهیم کرد تا در نهایت یک ارزیابی عادلانه از همه دانشجویان عزیز داشته باشیم. ما از Moss برای شناسایی فایل‌های مشابه استفاده خواهیم کرد.

۳. در صورت شناسایی تقلبی که ٪۵۰ یا پایینتر از کار را شامل می‌شود، اعضای گروه یک کارت زرد دریافت کرده و نمره «۰.۵- * بارم تمرین ۲» به اعضای گروه تعلق می‌گیرد و در صورت شناسایی تقلبی که بیشتر از ٪۵۰ کار را پوشش می‌دهد به اعضای گروه کارت قرمز تعلق گرفته و نمره «۱- * بارم تمرین ۲» به اعضای گروه تعلق می‌گیرد. علاوه بر این نمره منفی، گرفتن چند کارت زرد یا قرمز، عواقب شدیدتری خواهد داشت.

*در نهایت، هر گونه سوال در مورد تمرین و بخش‌های آنها را **تنها و تنها** از طریق سایت درس و ایجاد مباحثه با عناوین مرتبط مطرح بفرمایید.*

تندرست و موفق باشید

تیم درس مبانی رایانش ابری