به نام خدا

رایانش ابری

تمرین سوم

آشنایی با Hadoop و Spark (فاز دو)

طراحان تمرین:

محمدصادق محمدی، محمد رحمانیان

استاد درس:

دکتر جوادی

مهلت نهایی ارسال پاسخ:

۱۱ خرداد ساعت ۲۳:۵۹

مقدمه:

در بخش دوم شما با **Spark** آشنا خواهید شد.Apache Spark یک فریمورک نرم‌افزاری متن‌باز و چند منظوره است که قابلیت پردازش داده‌های بزرگ را در سرعت بالا با استفاده از محاسبات موازی فراهم می‌کند. Spark نسبت به Hadoop MapReduce، که از مدل MapReduce برای پردازش داده‌ها استفاده می‌کند و داده‌ها را در دیسک ذخیره می‌کند، عملکردی تا 100 برابر سریع‌تر ارائه می‌دهد، زیرا داده‌ها را در حافظه پردازش می‌کند که چرخه‌های پردازش داده را به‌طور قابل‌توجهی سرعت می‌بخشد. اسپارک از مدل RDD استفاده می‌کند که امکان پردازش موازی و یکپارچه‌سازی را فراهم می‌آورد تا عملکرد بهینه‌ای داشته باشد. این پلتفرم به دلیل عملکرد بالا، قابلیت پردازش در زمان واقعی، پشتیبانی از مجموعه‌ای از الگوریتم‌ها و کتابخانه‌های مختلف (برای یادگیری ماشین و تحلیل داده) و امکانات موازی‌سازی، بسیار محبوب است و می‌تواند بر روی یک کلاستر یا حتی یک محیط ابری اجرا شود.

در فاز دوم تمرین سوم شما باید برای از کلاستر فاز صفر استفاده کنید و وظایف مختلف را در Apache Spark انجام دهید.

آماده‌سازی برای تمرین:

برای اینکار لازم است که کلاستر خود را یک بار پایین آورده و دوباره آن را بالا بیاورید. (توجه داشته باشید که نیازی به پاک کردن ایمیج‌ها یا کار اضافه‌ای نیست)

۱- ابتدا به دایرکتوری که در فاز صفر از گیتهاب کلون کرده بودید بروید.
۲- دستور زیر را اجرا کنید تا تغییرات جدید دانلود شوند.

```
git pull
```

۳- کلاستر خود را با دستور زیر پایین بیاورید.

```
bash master-delete.sh
```

۴- دوباره کلاستر خود را بالا بیاورید.

```
bash master-build.sh
```

۵- داخل مرورگر خود localhost:8888 را وارد کرده تا UI نوتبوک را ببینید.

توجه داشته باشید که در قسمت آماده‌سازی لازم است تا برخی ایمیج‌ها و قسمت‌های دیگر دوباره دانلود شوند پس لازم است از DNS یا تحریم‌شکن مناسب استفاده کنید.

شرح تمرین:

این تکلیف طراحی شده است تا شما را با جنبه‌های عملی کار با Apache Spark آشنا کند و بر عملکرد برتر آن نسبت به فناوری‌های کلان داده سنتی مانند Hadoop، به ویژه از نظر سرعت به دلیل پردازش داده‌های درون حافظه، تأکید دارد. در طول این تمرین، شما درگیر وظایف مختلفی خواهید بود که در بخش های زیر هستند:

- راه اندازی و کانفیگ spark session
- Load و تغییر داده ها با استفاده از Spark DataFrames
- استفاده از Spark SQL برای انجام پرس‌وجوهای SQL برای تجزیه و تحلیل داده‌ها در Spark
- استفاده از RDD یا Resilient Distributed Datasets
- استفاده از user-defined functions برای گسترش قابلیت های اسپارک

هدف این وظایف در مجموع ارائه یک درک جامع از قابلیت‌ها و ابزارهای Spark است و شما را برای پردازش کارآمد مجموعه داده‌های بزرگ و انجام وظایف تجزیه و تحلیل داده‌ها آماده می‌کند.

بخش امتیازی:

- بهبود عملکرد در عملیات RDD را از طریق caching با مقایسه زمان‌های اجرا برای دسترسی به داده‌ها قبل و بعد از اعمال cache نشان دهید. توضیحات مربوط به نحوه بهبود را در گزارش خود ذکر کنید.
- مفاهیم narrow and wide transformations در اسپارک را توضیح دهید. چه چیزی یک narrow transformation را از wide transformation متمایز می کند؟ با استفاده نوتبوک Jupyter عملیات narrow and wide transformations را نشان دهید. زمان اجرای آن ها را اندازه گیری و مقایسه کنید.

گزارش:

به سوالات زیر در گزارش مربوط به این فاز پاسخ دهید:

- تفاوت های معماری و عملکردی بین RDDs و DataFrames در اسپارک را توضیح دهید. موارد استفاده مربوطه، جنبه های عملکرد و سطوح انتزاع را مورد بحث قرار دهید.
- Spark SQL را با DataFrame API در Apache Spark مقایسه کنید. تفاوت‌ها را از نظر قابلیت‌های پرس و جو، بهینه‌سازی عملکرد، و قابلیت استفاده API توضیح دهید. مثال‌هایی ارائه کنید که ممکن است یکی بر دیگری ترجیح داده شود.
- مفهوم partitioning داده را در اسپارک توضیح دهید. چرا partitioning در پردازش داده توزیع شده اهمیت دارد.
- استراتژی‌های partitioning مختلف را بررسی کنید و توضیح دهید چطور partitioning می‌تواند در عملکرد اجرای job در اسپارک تاثیر بگذارد.

نکات مربوط به تمرین تحویلی:

- تمرین شما تحویل اسکایپی خواهد داشت؛ بنابراین از استفاده از کدهای یکدیگر یا کدهای موجود در وب که قادر به توضیح داده عملکرد آنها نیستید، بپرهیزید.
- در صورت داشتن هرگونه مشکل، سوالی یا ابهام، آن را در با تدریس یاران درس مطرح کنید تا آن‌ها در سریع‌ترین زمان ممکن به شما پاسخ دهند.

مواردی که باید ارسال شود:

- یک فایل زیپ با نام studentID_HW3_0.zip که شامل گزارش شما به همراه نوتبوک شما است.

موفق باشید

تیم تدریس‌یاری مبانی رایانش ابری