

به نام خدا



## رایانش ابری

# آشنایی و کار با Hadoop و Map-Reduce

استاد درس:

خانم دکتر نورانی

طراحان پروژه:

آقایان محمدصدرا حائری اسدی، محمدرضا درودیان، آریان بوکانی و سیدعلیرضا اعزاز

مهلت نهایی ارسال پاسخ:

فاز ۱: نصب ماشین مجازی - ۵ آذر

فاز ۲: نصب هادوپ - ۱۲ آذر

فاز ۳: پیاده‌سازی مپ ردیوس - ۲۰ آذر

## قدم اول: نصب ماشین مجازی

برای پیاده‌سازی قدم‌های بعدی شما نیازمند نصب ماشین مجازی لینوکس هستید و به همین منظور نیاز به ۳ ماشین مجازی داریم.

به ماشین مجازی اول 1 vCPU و 1 GB Ram و 20 GB حافظه دیسک و به ماشین‌های مجازی دوم و سوم 2 vCPU و حافظه بیشتر مثلاً 2 GB Ram اختصاص دهید.

## قدم دوم: نصب و راه اندازی خوشه Hadoop

در این تمرین، یک خوشه Hadoop را با استفاده از سه ماشین مجازی راه اندازی و برنامه MapReduce را بر روی آن اجرا می‌کنید.

برای ایجاد ماشین‌های مجازی، نصب Hadoop و راه اندازی خوشه مراحل ذکر شده در [این لینک](#) را به دقت دنبال کنید.

### حتماً نکات زیر در مراحل پیاده‌سازی اعمال نمایید.

- در مرحله ۸ در لینک فوق، از لینک زیر به جای لینک ذکر شده در دستور استفاده نمایید:

<https://archive.apache.org/dist/hadoop/common/hadoop-3.2.2/hadoop-3.2.2.tar.gz>

- در مرحله ۹ و ۱۱ فقط کسانی که از سری پردازنده‌هایی با معماری arm (مانند Apple silicon یا همان پردازنده‌های سری M) استفاده می‌کنند، باید به جای amd، از arm در دستورات استفاده کنند.

- در مرحله ۱۹، مقدار replication را برابر 1 قرار دهید.

لطفاً به نکات زیر توجه داشته باشید:

- اگر مراحل را به درستی طی کرده باشید، نصب به گونه ای انجام می‌شود که ماشین مجازی اول نقش‌های NameNode و ResourceManager و ماشین‌های دوم و سوم نقش DataNode و NodeManager را به عهده خواهند گرفت. **با استفاده از دستور jps از صحت این مساله اطمینان حاصل کنید و از آن اسکرین‌شات تهیه کنید و در گزارش خود قرار دهید.**

- با قرار دادن اسکرین‌شات نشان دهید که WebGUI از کامپیوتر شخصی شما قابل دسترسی است.

## توضیحات مربوط به Dataset

- دیتاست داده شده، شامل 1.55 میلیون رکورد در مورد فیلم‌ها و سریال‌های IMDb می‌باشد.
- هر کدام از رکوردهای این فایل شامل ۹ ستون متفاوت هستند.
- از [این لینک](#) می‌توانید دیتاست را دانلود نمایید.
- اطلاعات مربوط به هر ستون دیتاست در پایین آورده شده‌اند:
  - ستون اول (tconst) : آیدی مربوط به هر رکورد
  - ستون دوم (titleType) : نوع (فیلم، سریال، ویدئو و ...) هر عنوان
  - ستون سوم (primaryTitle) : نامی که عنوان با آن شناخته شده است
  - ستون چهارم (originalTitle) : نام اصلی عنوان با زبان اصلی
  - ستون پنجم (isAdult) : صفر به معنی اینکه این عنوان برای عموم می‌باشد و ۱ به معنی اینکه تنها مناسب بزرگسالان می‌باشد
  - ستون ششم (startYear) : تاریخ عرضه‌ی عنوان
  - ستون هفتم (endYear) : تاریخ اتمام عنوان در صورتی که عنوان از نوع سریال باشد (برای بقیه‌ی عناوین موجود در دیتاست این مقدار \N می‌باشد)
  - ستون هشتم (runtimeMinutes) : طول عنوان به دقیقه
  - ستون نهم (genres) : ژانرهای عنوان

## قدم سوم: توسعه و اجرای برنامه MapReduce

1. با استفاده از HDFS CLI، پوشه‌ی /user/hadoop را در HDFS بسازید.
  2. پوشه‌ی dataset.zip را از لینک داده شده دانلود کرده و از حالت فشرده در بیاورید.
  3. فایل اکسترکت شده را در مسیر /user/hadoop/input قرار دهید.
  4. یک برنامه‌ی MapReduce ساده بنویسید که تعداد هر کدام از برنامه‌های (titleType) movie, short, tvSeries و tvEpisode را حساب کند. طبیعتاً خروجی باید برای هر کدام از این چهار دسته عددی را برگرداند.
- توجه:** فایل خروجی نباید اطلاعات دیگری را شامل شود.

5. یک برنامه MapReduce بنویسید که نشان دهد چه تعداد از برنامه‌ها دارای primaryTitle و originalTitle یکسانی هستند، titleType آنها **movie** یا **tvSeries** باشد و جزو فیلم‌های بزرگسالان باشد یعنی isAdult آن برابر با 1 باشد. در واقع خروجی شما باید تعداد برنامه‌ها را با این ویژگی‌های ذکر شده (یعنی دو نوع title با هم برابر باشند و فیلم بزرگسال باشد) برای دو دسته برنامه **movie** و **tvSeries** بدهد. توجه: فایل خروجی نباید اطلاعات دیگری را شامل شود.

	counts
movie	?
tvseries	?

6. یک برنامه MapReduce بنویسید که نشان دهد چه تعداد از برنامه‌ها دارای titleType برابر با **tvEpisode** یا **tvSeries** اند و سال شروع آنها (startYear) بزرگتر مساوی ۱۹۹۹ و کوچکتر مساوی ۲۰۱۵ است و نباید ژانر (genres) آن **Biography** باشد. در ضمن اگر سال پایان نداشت، نباید آن رکورد نمایش داده شود.

## نکات مربوط به تحویل تمرین

- برای راهنمایی می‌توانید از ویدئوهای تهیه شده توسط تدریس‌یاران کمک بگیرید.
- تمرین شما در فاز ۳ تحویل اسکایپی خواهد داشت. بنابراین از استفاده از کدهای یکدیگر یا کدهای موجود در وب که قادر به توضیح داده عملکرد آنها نیستید، بپرهیزید!
- ابهامات خود را به ایمیل تدریس‌یاری ارسال کرده و ما در سریع‌ترین زمان ممکن به آنها پاسخ خواهیم داد.
- ایمیل: [cciust01@gmail.com](mailto:cciust01@gmail.com)

## **آنچه که باید ارسال کنید**

- **برای فاز ۱ :** اسکرین‌شات از مراحل نصب و راه‌اندازی ماشین مجازی (داخل یک فایل pdf با نام StudentID\_part1.pdf قرار دهید.)
- **برای فاز ۲ :** اسکرین‌شات از مراحل نصب و راه‌اندازی هدوپ (داخل یک فایل pdf با نام StudentID\_part2.pdf قرار دهید.)
- **برای فاز ۳:** فایل مربوط به کدهای MapReduce و فایل‌های نتایج (داخل یک فایل zip با نام StudentID\_part3.zip قرار دهید.)

**موفق باشید**

**تیم درس رایانش ابری**