Hail with Amazon EMR notebook

??? ??

1. EMR 스튜디오에서 동일한 VPC에 새 스튜디오를 만듭니다 .
2. 생성한 스튜디오에 대한 새 작업 공간 (노트북)을 생성하면 완료되면 jupyterHub에 대한 새 탭이 자동으로 열립니다 .
3. python3 커널을 사용하여 새 노트북을 시작합니다 .

EMR ???? Hail ??

```
pip install hail
```

추후 **S3a** 프로토콜 사용을 위한 내용 (Hadoop-AWS module)

1. Ssh into primary node (**as sudo user**)
2. Go to the jars directory: `cd /home/emr-notebook/.local/lib/python3.9/site-packages/pyspark/jars`
3. Download the 2 jar files with the following command in the directory:
   1. sudo curl -sSL https://search.maven.org/remotecontent?filepath=org/apache/hadoop/hadoop-aws/3.3.2/hadoop-aws-3.3.2.jar > ./hadoop-aws-3.3.2.jar
   2. sudo curl -sSL https://search.maven.org/remotecontent?filepath=com/amazonaws/aws-java-sdk-bundle/1.12.99/aws-java-sdk-bundle-1.12.99.jar > ./aws-java-sdk-bundle-1.12.99.jar

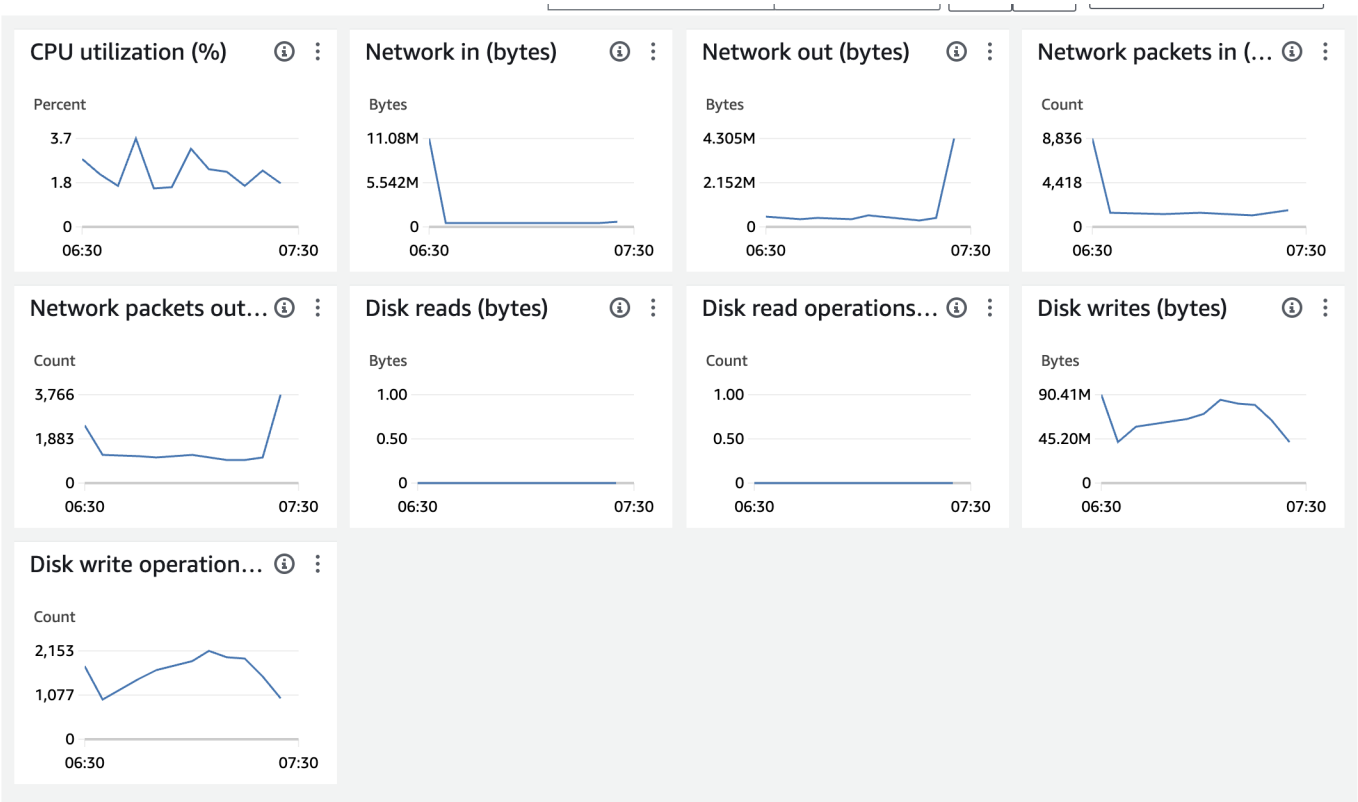
의존성 설치

https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.6/

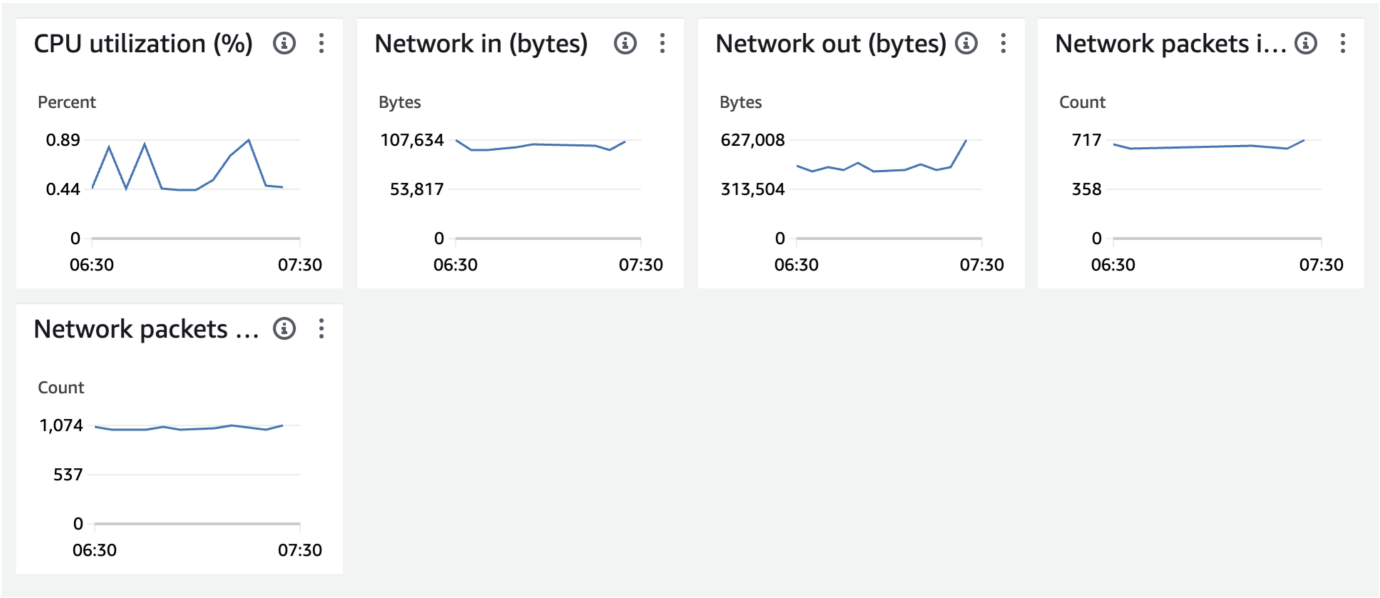
https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/

There are 2 jars missing in the java class path the notebook is using. Using python shell directly from the cluster does not need to do this (but only need to point `SPARK_HOME` to the jars because the required dependencies are already there if run from hadoop environment) as Notebook hosts a different environment for all dependencies installed. Also, the hadoop version (uses aws version of hadoop) and package is slightly different from what we get in the hadoop environment in `SPARK_HOME;` Notebook environment uses the external hadoop client, meaning that it will not be able to connect to S3.
We need to download aws sdk jar and hadoop aws jar and put them into Notebook's environment jar collection.

Primary □□□ □□□□ EC2 Monitoring



Core □□□ □□□□ EC2 Monitoring



예제 노트북 다운로드

hail-tutorial.zip

????

- https://hail.is/docs/0.2/tutorials/01-genome-wide-association-study.html#Quality-Control
- https://github.com/hmkim/quickstart-hail/tree/main/packer-files/scripts
- **EMR on EC2**로 **Hail** 쥬피터를 사용하기 위한 **EMR** 노트북 환경 구성
  - https://catalog.us-east-1.prod.workshops.aws/workshops/c86bd131-f6bf-4e8f-b798-58fd450d3c44/en-US/emr-notebooks-sagemaker
- **EMR Serverless** 로 **Hail** 작업제출하기
  - https://catalog.us-east-1.prod.workshops.aws/workshops/f9855d43-62e3-441b-ba02-7f37a278c077/en-US/5-emr-serverless

---

Revision #3
Created 3 May 2024 15:43:04 by Hyunmin Kim
Updated 21 May 2024 07:59:27 by Hyunmin Kim