

Hail with Amazon EMR notebook

??? ??

1. EMR `elasticmapreduce` `aws` VPC `aws` `aws` .
2. `aws` `elasticmapreduce` `aws` `aws` `aws` (`aws`) `aws` `aws` `jupyterHub` `aws` `aws` `aws` .
3. `python3` `aws` `aws` `aws` `aws` .

EMR ???? Hail ??

```
pip install hail
```

`aws` **S3a** `aws` `aws` `aws` (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`

`aws`

`aws`

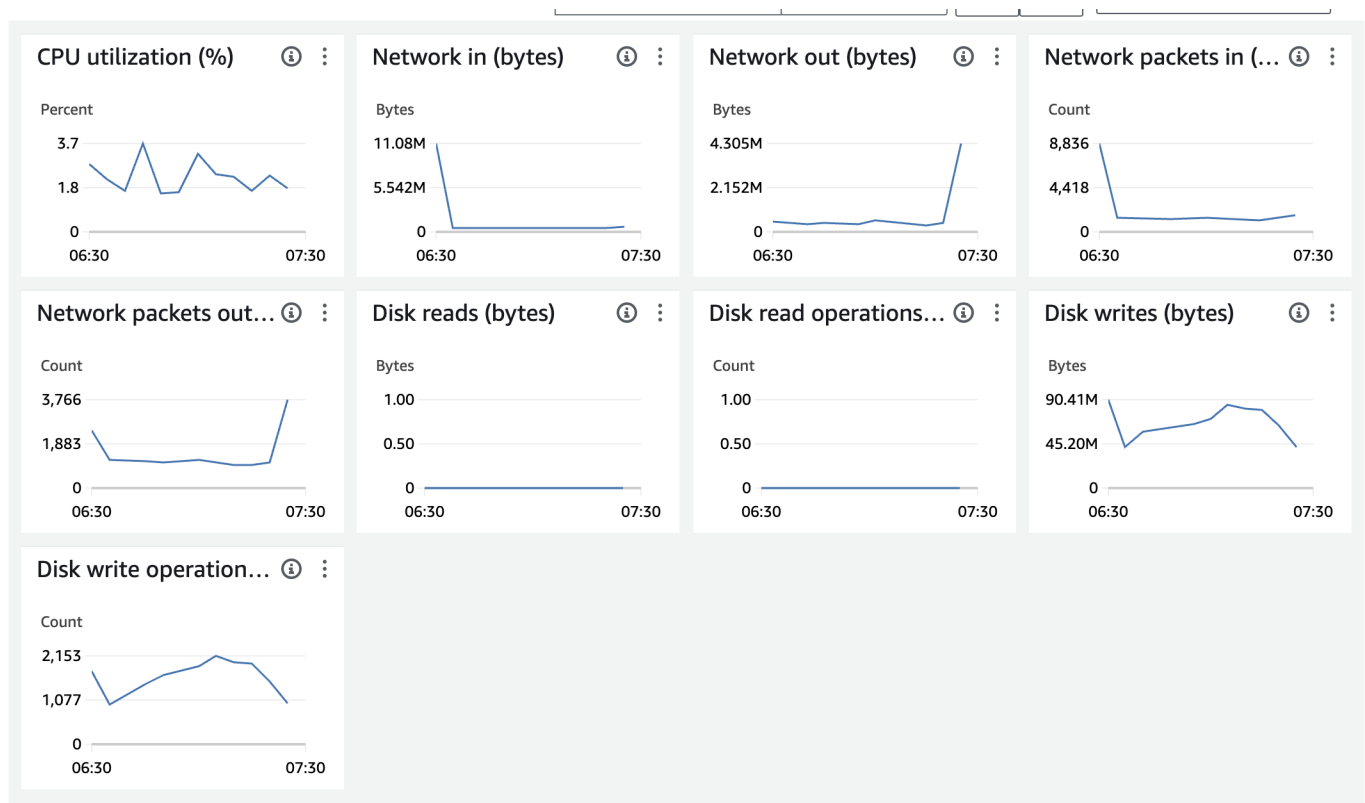
<https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.6/>

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

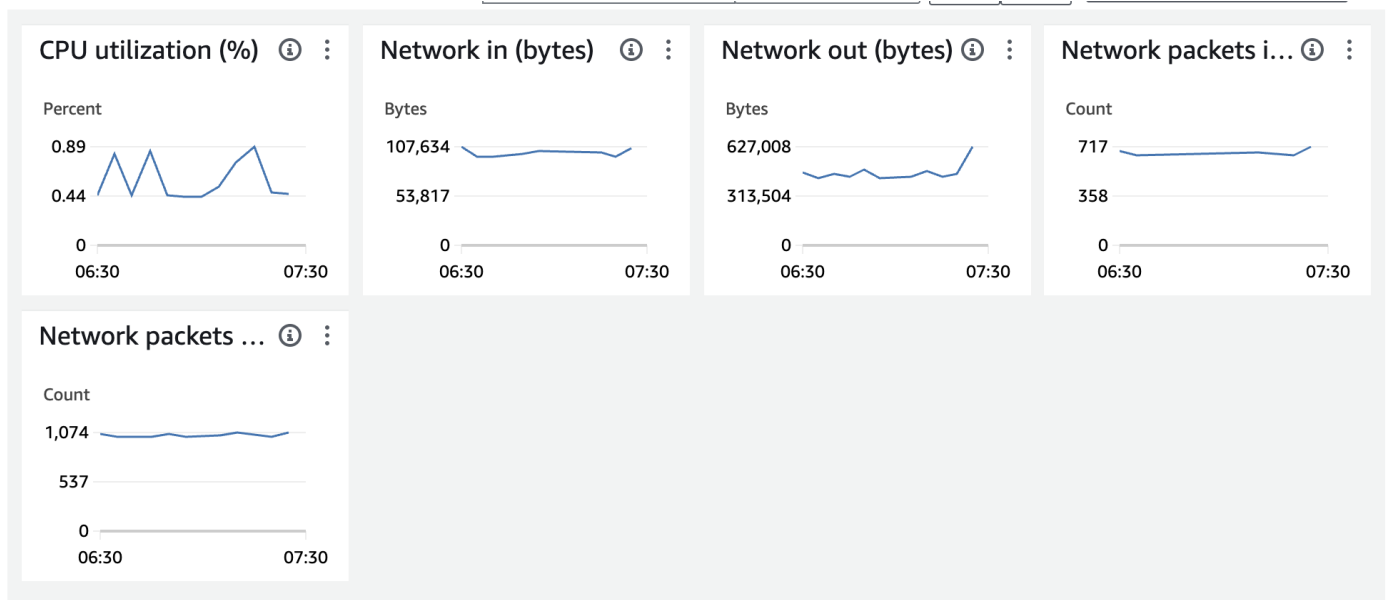
There are 2 jars missing in the java class path the notebook is using. Using python shell directly from the cluster does not need to do this (but only need to point `SPARK_HOME` to the jars because the required dependencies are already there if run from hadoop environment) as Notebook hosts a different 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



□□ □□□ □□□□

[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** □ **Hail** □□□ □□□ □ **EMR** □□ □ □
 - <https://catalog.us-east-1.prod.workshops.aws/workshops/c86bd131-f6bf-4e8f-b798-58fd450d3c44/en-US/emr-notebooks-sagemaker>
- **EMR Serverless** □ **Hail** □□□□□
 - <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