Managing “Cluster Top” resources

Spark + rapids.ai

Spark

- Supports Data Science and Machine Learning
- Several nodes concurrently
- Hadoop model
  - Efficient parallelism for data manipulation

Spark

- Simplified installation
- Can be used from API (like Pandas)
  - Pyspark
- Submitting job to a cluster
  - spark-submit
- Do not support GPU (default)

Spark - Monitoring

Spark Master at spark://10.46.0.14:8893

URL: spark://10.46.0.14:8893
Alive Workers: 5
Cores in use: 312 Total, 0 Used
Memory in use: 528.4 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 9 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (5)

Worker Id	Address	State	Cores	Memory	Resources
worker-20210414080220-10.46.0.11-46761	10.46.0.11:46761	ALIVE	32 (0 Used)	61.8 GiB (0.0 B Used)	
worker-20210414090314-200.145.46.221-36233	200.145.46.221:36233	ALIVE	96 (0 Used)	186.6 GiB (0.0 B Used)	
worker-20210414113941-10.46.0.13-32837	10.46.0.13:32837	ALIVE	72 (0 Used)	124.8 GiB (0.0 B Used)	
worker-20210414114331-10.46.0.12-40967	10.46.0.12:40967	ALIVE	72 (0 Used)	124.8 GiB (0.0 B Used)	
worker-20210414124012-10.46.0.15-39803	10.46.0.15:39803	ALIVE	40 (0 Used)	30.4 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (9)

How to explore GPUs?

1) Nvidia support execution ot spark jobs with rapids.ai

2) Dask

Execute machine learning on GPU

Utilization of multiple nodes and GPUs

Suggestions

Install spark for all users

Investigate the use of GPUs in spark or dask