Spark

Krótki opis usługi

Moduł umożliwia korzystanie z najpopularniejszych narzędzi przetwarzania danych typu BigData uruchamianych na zasobach Infrastruktury PLGrid. Obecnie moduł jest dostępna na klastrze Zeus i Prometeusz, gdzie można prowadzić obliczenia wielowęzłowe z wykorzystaniem systemów Spark lub Hadoop. W celu efektywnego używania oprogramowania Spark i Hadoop zalecamy zapoznanie się z Spark Programming Guide.

Aktywowanie usługi

Dostępne w ramach modułu plgrid/apps/spark

Pierwsze kroki

Uruchomienie zadania Spark wykorzystującego 4 rdzenie na 1 węźle w trybie klastra Spark (Spark Standalone cluster in client deploy mode):

$ srun -N1 --ntasks-per-node=4 --pty /bin/bash

Wykorzystanie zostanie zaliczone na konto grantu osobistego PLGrid. W przypadku potrzeby podania innego grantu należy wykorzystać opcję -A specyfikując identyfikator aktywnego grantu PLGrid:

```
$ module load plgrid/apps/spark
```

$ start_spark_cluster

Obliczenia:

```
$ $SPARK_HOME/bin/spark-submit $SPARK_HOME/examples/src/main/python/wordcount.py /etc/passwd
```

Zatrzymywanie klastra:

$ stop_spark_cluster

Uwagi

1. Instalacja w ACK CYFRONET nie udostępnia systemu plików HDFS, z tego powodu korzystanie z komend 'hdfs' oraz 'hadoop fs/dfs' jest niemożliwe.
2. Na wybranym węźle może być uruchomiony tylko jeden Master program w tym samym czasie. W przypadku próby uruchomienia kolejnego klastra BigData na tym samym węźle, zostanie wyświetlony komunikat, aby spróbować na innym węźle.

Zaawansowane użycie

Dokumentacja Hadoop oraz Spark w Cyfronet

Gdzie szukać dalszych informacji?

Oficjalny Spark Programming Guide