Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

GPU használata Big Data felderítő adatanalízise során

Furó János Olivér, IV. évf, (BSc) mérnök inf. szakos hallgató
Konzulens: Kocsis Imre, MIT
Informatikai technológiák szakirány, rendszertervezés ágazat
Önálló laboratórium összefoglaló
2015/16. II. félév

A félév során megvizsgáltam, hogy hogyan lehet kihasználni a grafikus processzorok nyújtotta számítási teljesítményt Big Data felderítő adatanalízise során. Legfőbb kérdés az volt, hogy mekkora gyorsulást lehet elérni így a két terület ötvözésével. Az adathalmaz méretére való megkötés annyi volt, hogy férjen el a számítógép, illetve a grafikus eszköz memóriájában.

Megismerkedtem a felderítő adatanalízissel és az adatok grafikus megjelenítésével valamint azzal, hogy milyen kihívások jelennek meg az elemzésre kerülő adathalmaz méretének növekedésével.

Kiindulási alapnak Hadley Wickham bin-summarise-smooth keretrendszerét vettük, ami egy R-hez elérhető csomag Bigvis néven és megalkotásának célja a felderítő adatanalízis felgyorsítása volt. Ez jelenleg az egyik legjobb megoldás nagy adathalmazok (nagyságrendileg akár 100 millió megfigyelés) vizualizálásához, maga mögé utasítva a Mondrian-t és ggplot2-t.

A célunk az volt, hogy gyorsabb megjelenítést érjünk el, mint amire a bigvis képes, mégpedig azzal, hogy a bin-summarise-smooth keretrendszer alapjait megvalósítjuk GPU-n.

A grafikus processzorok általános célú programozására az Nvidia által szolgáltatott CUDA C/C++ platformot használtam.

A megvalósítás során a számításigényes műveleteket a GPU végezte, egy C++ programba ágyazva. Ebben a programban történt az adatok beolvasása, eredmények kiírása, valamint a gyorsulás mérésének alapjául szolgáló műveletek elvégzése CPU-n és különböző futási idők mérése. Az eredmény megjelenítéséhez R-t és ggplot2-t használtam.

A mérések és tesztek elvégzése egy átlagosnak mondható laptopon történt, CUDA tekintetében belépő szintű eszközön.

Az eredmények azt mutatták, hogy a tényleges számítási műveleteken a GPU használatával kedvező esetben akár 12x-es gyorsulás is elérhető (CPU-hoz képest), azonban ez nagyban függ a bemeneti adathalmaztól és annak rendezettségétől.

A mérések során különböző adathalmazokat és különböző megoldásokat is vizsgáltam. Ahhoz hogy ne csak kedvező esetekben lehessen elérni a gyorsulást és ahhoz, hogy ennél jobbat is sikerüljön, további elmélyedés szükséges a témában. A kezdeti sikerek fényében ezt mindenképpen érdemes folytatni és tovább gondolni, akár interaktív vizualizáció megvalósításának tekintetében.