Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

Nagyméretű adatsorok statikus vizualizációja

Bereczki Ádám, (BSc) mérnök informatikus szakos hallgató
Konzulens: Salánki Ágnes, MIT
Informatikai technológiák szakirány, Rendszertervezés ágazat
Önálló laboratórium összefoglaló
2015/16. II. félév

Napjainkban az informatikai technológiák fejlődése és alkalmazása során egyre gyakrabban van szükség nagyméretű adathalmazok feldolgozására, egy adott rendszer fejlesztése és üzemeltetése során. Az önálló laboratórium során a feladatom az efféle adatsorok felhasználásának kezdeti felderítő analízis fázisában jelentős szerepet kapó vizualizáció megismerése és alkalmazása volt.

A célom a teljesítmény megfigyelésekből származó adatokból összetett diagrammokon vizualizálni a rendszer állapotát különböző terheltségű időszakokban. Ezen felül a kapott diagrammok elemzése során tapasztalt kiugró, az átlagtól nagyban eltérő események okára való következtetés a foglalási adatokból.

A munkám során a tanszéki Virtual Computing Lab (VCL) cloud rendszer adatait használtam. Az adatok a VLC rendszer tíz darab hosztjának teljesítményadatait, és a rendszerhez tartozó foglalási adatokat foglalta magába.

A feladatmegoldást megelőző szakirodalmi kutatás során megismerkedtem a Hadley Wickham Bin-Summarise-Smooth[1] írásával, melyben egy hatékony módszert mutat be nagyméretű adatok vizualizációjára. A módszer lényege, hogy elkülönítjük a számításokat az adatok megjelenítésétől. A megoldásom során ezt a módszert követtem.

A feladat megvalósításához az R nyelvet választottam. Ez a nyelv kifejezetten statisztikai számítások elvégzésére lett tervezve, ezért döntöttem a használata mellett.

Az adatok mennyisége miatt azok mérete meghaladta a rendelkezésre álló memória méretét ezért az adatok feldolgozása több részletben történt. Az adatsorok darabolását és a memóriába mozgatást a `data.table` nevű R package használatával valósítottam meg. Az beolvasáshoz használt eszköz kiválasztásakor több függvényt is megvizsgáltam (read.csv, fread, read.big.matrix, read.csv.ffdf) és a sebesség alapján döntöttem a `data.table` fread függvénye mellett.

A számítási rész tervezésekor megvizsgáltam a H2o nevű, nyílt forrású, bigdata elemzéshez használt eszköz R nyelvhez kiadott csomagját, felhasználásának elvi lehetőségét. A H2O eredetileg predikciók meghatározására és az adatban lévő minták hatékony felismerésére van tervezve. Arra voltam kíváncsi, hogy a gépi tanuláshoz implementált függvények alkalmasak-e vizualizációs feladatokhoz használt számítások elvégzésére, így a használata mellett döntöttem.

A grafikus megjelenítéshez a Leland Wilkinson vizualizációs nyelvének a Grammar of Graphics-nek a Hadley Wickham féle implementációját a ggplot2 adatvizualizációs csomagját használtam. Ezt a könyvtárat ideálisnak találtam a feldolgozott adatok minőségi megjelenítéséhez.

A labor végére sikeresen ábrázoltam a merevlemez írási és olvasási sebességadatait, a hálózati terhelés adatait kimenő és bejövő forgalom esetén, valamint a processzor és memória terheltséget, és esetenként megvizsgáltam, hogy az adott processzor és memória terheltségi szinteket mely hosztok szolgáltatják.

A továbbiakban sor kerülne a kiugró értékek automatikus megtalálására és ezek vizsgálatára a foglalási adatok segítségével. Megnézném, hogy a nagyobb teljesítményingadozások milyen események hatására következnek be.

A feladat végén a H2O-t nem tartom hatékony megoldásnak, bizonyos részei hiányosak az ilyen típusú felhasználáshoz.

1 http://vita.had.co.nz/papers/bigvis.pdf