Feladatkiírás – Elosztott gráftároló készítése

Integrált fejlesztés Java platformon (BMEVIIIJV51) házi feladat specifikációja.

- Név: Szárnyas Gábor
- Neptun kód: U944EQ

Feladat rövid szöveges ismertetése

Nagyméretű gráfok tárolásánál felmerülő igény, hogy a gráfot egy elosztott, skálázható rendszerben helyezzük el. A jelenleg elérhető nyílt forráskódú technológiák közül néhány már lehetővé teszi az elosztott tárolást, ilyen pl. a Titan keretrendszer (http://thinkaurelius.github.com/titan/). A Titan keretrendszerrel a gráfot GraphML (http://graphml.graphdrawing.org/) formátumból tölthetjük be. A keretrendszer a gráfot egy elosztott NoSQL adatbáziskezelőkből álló fürtre képzi le (jelenleg az HBase és a Cassandra NoSQL rendszerek támogatottak).

Bár a Titan támogatja a konkurens hozzáférést, a GraphML állomány betöltése csak egy Titan csomóponton keresztül valósul meg, így nem párhuzamosítható. A feladat célja egy elosztott gráfbetöltő alkalmazás készítése, amely képes egy elosztott fájlrendszeren tárolt GraphML fájl párhuzamos feldolgozására és betöltésére egy Cassandra adatbáziskezelőkből álló fürtbe. A feladatnak nem célja, hogy a tárolt gráfon komplex lekérdezéseket lehessen futtatni vagy mintaillesztéseket lehessen végezni; a lekérdezések csupán a gráf adott típusú/azonosítójú csomópontjainak és éleinek listázására szorítkoznak. Az elkészült alkalmazást a diplomatervem keretében készített inkrementális mintaillesztőben fogom kipróbálni a Titan alternatívájaként.

Az elkészült alkalmazás a későbbiekben felhasználható arra, hogy gráfpartícionáló algoritmusok implementációját teszteljük és azok hatékonyságát megvizsgálhassuk.

Use case-ek felsorolása

- Cassandra fürt elindítása
- GraphML állományként tárolt gráf leképzés és betöltése a Cassandra fürtbe
- A gráf csomópontjainak és éleinek lekérdezése
- Azonosítójával megadott csomópontok és élek törlése a gráfból

Alkalmazni kívánt technológiák felsorolása

- Kötelező technológiák:
  - GUI: webes (HTML5, AJAX)
  - Elosztott architektúra: JAX-RS (REST)
  - Tesztelés: JUnit
  - Automatikus fordítás: Maven
  - Naplózás: Log4j
  - Szerializálás: Thrift (CQL3 mögött)
- Egyéb technológiák:
  - Aszinkron kommunikáció: Akka
  - Elosztott fájlrendszer: HDFS
  - Elosztott adatbázis-kezelő: Cassandra