# Twitter adatok adatbázisba töltése

## Az adatbázisok

## Az adatok

A twitter adatok két különböző „stream”-en jönnek. Az egyik a gyakran „world stream”-nek nevezett adatfolyam, ami nagyrészt GPS koordinátás üzeneteket tartalmaz, de ebben elég kevés a retweet. A másik a „retweet stream”, amiben mindenféle üzenet van vegyesen. A két adatfolyam nagyjából diszjunk. Két eszközzel lehet az adatokat lementeni. Egyrészt a curl nevű programmal, ami egyszerűen elmenti a nyers adatokat egy fájlba, illetve van a TwitterToolkit nevű saját fejlesztésű program, ami feldarabolva, tömörítve menti el az adatfolyamot. Ez a feldarabolás nagyon lényeges, ugyanis az adatbázisba írás „chunk”-okban történik. Ezek optimális mérete 10GB körül van. Ezért a curl által lementett nyers streameket is fel kell darabolni adatbázisba töltés előtt. Másrészt a nyers stream mentéseket eléggé nagy helypazarlás tömörítés nélkül tárolni.

Az adatokat ún. „run”-okba szervezvezzük, és úgy töltjük be az adatbázisba. A run méretére nincsen korlát, igazából szubjektív, hogy mit nevezünk runnak.

## A TwitterToolkit

A TwitterToolkit egy több programból álló csomag, amivel az adatok SQL Serverbe tölthetők. A programcsomag a <https://svn.vo.elte.hu/svn/vo/dobos/TwitterToolkit> repositoryból tölthető le, és VS 2010-zel fordítható. A főbb komponensei: twread, twload, twbatch.

## A twread program

Ez a program az élő twitter streamek olvasására, lemezre mentésére, illetve már lementett nyers streamek feldarabolására és tömörítésére való. Használata:

**twread sample -compress -output <path> -linemax <numlines>**

Ez a parancssor tömörítve lementi a twitter sample streamet (kb. 0.5%-a az összes üzenetnek, tartalmaz retweeteket. Megadható, hogy hány soronként kezdődjön új fájl. Alternatív megoldásként megadható a -timemax <minutes> paraméter, ami meghatározza, hogy mennyi időnként kezdődjön új fájt. Napi egy fájl írása általában célravezető (1440 perc).

**twread worls -compress -output <path> -linemax <numlines>**

Ez a parancssor az előzőhöz hasonlóan működik, de a GPS koordinátás üzeneteket menti le.

twread chunk -compress -input <infile> -output <path> -timemax <time>

Ez szintén a fentiekhez hasonlóan működik, de nem az elő twitter streamet olvassa, hanem egy már, pl. curl használatával lementett fájlból olvas, és abból készít kisebb méretű chunkokat, amik már alkalmasak adatbázisba töltésre.

A program jelenleg http authentikációt tud, a felhasználónév és a jelszó a twread.exe.config fájlban állítható be.

## A twload program

Az adatbázisba töltés több lépésből áll. A nyers twitter adatok formátuma JSON, ami egy javascript számára könnyen olvasható szöveges formátum. Ezt adatbázisba írás előtt át kell alakítani egy olyan formátumra, amit az adatbázis-szerver megért. Ezek a fájlok lesznek az ún. „bulk load” fájlok, melyek lehetnek szövegesek vagy binárisak. Ezután az adatok egy átmeneti helyre, egy ún. „LoadDB”-be kerülnek, majd innen lehet őket „merge”-ölni a nagy adatbázisba.

Az éles adatbázison számos index van, ami a kereséseket igen meggyorsítja, viszont az adatok merge-ölését lehetetlenné teszi. A merge idejére ezért az indexeket ki kell kapcsolni. Ez azt jelenti, hogy az adatbázis gyakorlatilag használhatatlanná válik. A folytonosság fenntartása érdekében ezért hosszú távon két példány lesz az adatbázisból, és amikor az egyikbe már bekerültek az új adatok, akkor azt újraindexeljük, majd a két adatbázist megcseréljük, és a másikba is betöltjük az adatokat.

Az adatbázisba töltés pontos lépései:

1. prepare: ilyenkor a kiindulási JSON fájlt feldolgozzuk, és bulk insert fájlokat írunk. A bulk insert fájlokat az adatbázis-szerver már közvetlenül képes olvasni, és villám gyorsan feldolgozni. Fontos kitétel viszont, hogy a bulk insert fájloknak mindig az adatbázis-szerver lokális könyvtárában kell lenniük, mert hálózatról a szerver nem tudja őket betölteni. Ezért a bulk fájlokat tartalmazó könyvtárakra is mindig lokális, abszolút fájlnévvel kell hivatkozni, pl. c:\data\raid6\_0\temp…
2. load: ebben a lépésben a bulk insert fájlokat betöltjük egy átmeneti adatbázisba, a TwitterLoaderbe. Az adatok innen kerülnek majd át a végleges helyükre. Minden chunkhoz több tábla is generálódik a TwitterLoader adatbázisban.
3. merge: ekkor fésüljük össze a már betöltött adatokat az újonnan betöltendő chunkkal. Ez az összes közül a legerőforrásigényesebb művelet, kb. 4-5 percig tart egy napnyi adatra.
4. cleanup: kitakarítjuk a TwitterLoader adatbázis, és letöröljük a chunkokhoz kreált bulk insert fájlokat

A program rendelkezik egy batch paranccsal is, de ez elavult, helyette a twbatch program használandó.

## A twbatch program

Ez a betöltőprogram *legfrissebb verziója*, mely naplót készít a betöltés állapotáról, és a betöltéseket a megfelelő ponton tudja folytatni. Összesen két üzemmódja van:

1. create: ez egy bejegyzést készít a TwitterAdmin adatbázisba, melybe bekerül, hogy milyen adatokat akarunk betölteni. A betöltéshez manuális kell run\_id rendelni. A program eredményül visszaad egy batch\_id-t, amire hivatkozva elindítható, folytatható a betöltési művelet
2. start: ezzel a paranccsal indítható el egy batch betöltése, aminek a create művelet által visszaadott batch\_id-t kell megadni.

A TwitterAdmin adatbázis elérési adatai a twbatch.exe.config fájlban vannak, de ezt nem szükséges megváltoztatni. Minden batch esetében külön meg kell adni a cél-adatbázist, így ugyanaz a TwitterAdmin használható tetszőleges batch betöltésére, akár tesztelésre is.

A twbatch program a batch „start” paranccsal való elindítása után lekapcsolja a cél-adatbázison levő összes indexet, majd a batch végén újraépíti. A lekapcsolás egy pillanat, viszont a visszaépítés pár órát vesz igénybe.

**Példa:**

.\twbatch.exe create -runid 1003 -Source \\blackhole\staging\nmvo\twitter\runs\location \run\_1003\_nr\world3\_1.txt.gz -BulkPath C:\Data\Temp\nmvo\twitter\bulk -Target Twitter\_TEST -Server RETDB02 -E -Binary

-E jelentése: Windows authentikáció a szerveren

-Binary: bináris bulk insert fájlok használata

A fenti parancsra a program válasza:

Found 1 file(s), now creating batch...  
New batch created with ID 10

Ezután a betöltés elindítása:

.\twbatch.exe start -batchid 10

## Amire egyelőre nincsen kész program

Friend-follower gráf betöltő -> ez elvileg kész, de tesztelni kell

Koordináta klaszterezés -> a klaszterező kód készen van, a betöltőt tesztelni kell

Szabadszöveges indexelés -> erre csak félkész scriptek vannak, amit kézzel kell futtatni.