Análisis de grandes colecciones de datos con **Pig Latin**

El objetivo de este ejercicio es demostrar el uso del lenguaje Pig Latin para hacer *análisis de grandes colecciones de datos*. Para ello usted trabajará sobre una colección de datos reales: *resultados de pruebas de velocidad de carga y descarga* realizadas en diferentes lugares, momentos y con diferentes proveedores de red.[1]

Requisitos

- Pig latin 0.14.0 (o superior)
- Java 1.7 (o superior)
- Colección de datos Neubot (provista durante el curso)
- Archivos *neubot.pig* y *NeubotTestsUDFs.jar* (provistos durante el curso)

Esquema de los datos

La siguiente tabla enumera los datos más importantes obtenidos durante una prueba de velocidad:

Nombre	Descripción
client_address	Dirección IP del usuario (IPv4 o IPv6).
client_country	País en donde se realizo la prueba.
client_provider	Nombre del proveedor de red del usuario.
connect_time	Número de segundos que tomó el envío y recepción de paquetes durante la prueba (*Round-Trip Time*).
download_speed	Velocidad de descarga (medido en bytes/segs). Resultado de dividir el número de bytes recibidos entre el tiempo de descarga.
neubot_version	Versión del programa utilizado la prueba.
platform	Sistema operativo utilizado para la prueba.
remote_address	Dirección IP (IPv4 o IPv6) del servidor utilizado para la prueba.
test_name	Tipo de prueba (ej., *speedtest, bittorrent, dash*).
timestamp	Instante en el que la prueba fue realizada. Expresado en el número de segundos que han pasado a partir del 1/01/1970 (cf. estampilla UNIX).
upload_speed	Velocidad de carga. Medido como el número de bytes enviados durante el tiempo de envío.
latency	Numero de segundos que tardó el envío y recepción de un paquete de control (Round-Trip Time).
uuid	Identificador del usuario. Número generado aleatoriamente al momento de la instalación de Neubot).
asnum	Identificador alfanumérico del proveedor de red
region	Nombre de la región donde se realizo la prueba (si aplica).
city	Nombre de la ciudad.
hour	
month	Hora / mes / año de la prueba (derivado del timestamp).
year	

[1] http://neubot.org

Ejecución de Pig

Pig puede ser ejecutado en *modo local* o sobre una infraestructura Hadoop (i.e., un clúster de máquinas). El siguiente comando ilustra como correr un script pig en **modo local**:

```
/* Local mode */
$ pig -x local neubot.pig
```

El script llamado **neubot.pig** ilustra el uso de Pig Latin para cargar la colección de datos Neubot en memoria y realizar operaciones sobre dichos datos. En particular el script ejemplifica como:

- Cargar los datos y adaptarlos a un esquema particular.
- Filtrar las pruebas de velocidad de carga y descarga (i.e., speedtest).
- Conservar los nombres de las ciudades donde las pruebas fueron realizadas.
- Desplegar el resultado en pantalla.

```
REGISTER NeubotTestsUDFs.jar;
DEFINE   IPtoNumber convert.IpToNumber();
DEFINE   NumberToIP convert.NumberToIp();

NeubotTests = LOAD 'NeubotTests' using PigStorage(';') as (
            client_address: chararray,
            client_country: chararray,
            lon: float,
            lat: float,
            client_provider: chararray,
            mlabservername:  chararray,
            connect_time:    float,
            download_speed:  float,
            neubot_version:  float,
            platform:        chararray,
            remote_address:  chararray,
            test_name:       chararray,
            timestamp:       long,
            upload_speed:    float,
            latency:  float,
            uuid:     chararray,
            asnum:    chararray,
            region:   chararray,
            city:     chararray,
            hour:     int,
            month:    int,
            year:     int,
            weekday:  int,
            day:      int,
            filedate: chararray
);

--
```

```
-- Keep only the 'speedtests'
--
Tests = FILTER @ BY (test_name matches '.*speedtest.*');

--
-- Cities were the tests were conducted
--
Cities = FOREACH @ GENERATE city;
Cities = DISTINCT @;
Cities = ORDER @ BY city;
DUMP @;
```

Actividad a realizar

- Filtrar los speedtests que fueron hechos en 'Buenos Aires' o en 'Montevideo'. Después listar los proveedores en dichos países.
- Listar los nombres de los proveedores de red y los rangos de IPs con los que trabajan. Para esto use la función *IPtoNumber* definida en el script.
- Organizar las pruebas con respecto a si fueron hechas sobre 3G o en conexión filiar. Para esto haga la hipótesis de las velocidades máximas de descarga ofrecidas en cada tipo de conexión.
- Determinar cuál es el usuario que realizó el mayor número de pruebas. Para ese usuario en particular (i) obtenga el histórico de velocidades de descarga y muestre en una tabla la evolución de la velocidad de descarga con respecto al tiempo, (ii) enumere los países en los que el usuario ha realizado dichas pruebas.