Improving Online Social Network collection and processing mechanisms

Dimitris Tsagkarakis, Alexandros Vavakos, Vasilis Stavrou, Miltiadis Kandias

Information Security and Critical Infrastructure Protection Laboratory
Dept. of Informatics, Athens University of Economics & Business (AUEB)

Introduction
- Rapid explosion of Online Social Networks.
- Users transfer their offline behavior to the online world.
- Extraction of information from social networks contributes to the profiting of users.
- Open Source INTelligence (OSINT) to mitigate the insider threat.

<table>
<thead>
<tr>
<th>OLTP vs. OLAP</th>
<th>OLTP System</th>
<th>OLAP System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inserts and Updates</td>
<td>Short and fast inserts and updates initiated by end users</td>
<td>Periodic long-running batch jobs refresh the data</td>
</tr>
<tr>
<td>Queries</td>
<td>Relatively standardized and simple queries that return relatively few records</td>
<td>Often complex queries involving aggregations</td>
</tr>
<tr>
<td>Processing Speed</td>
<td>Typically very fast</td>
<td>Depends on the amount of data involved</td>
</tr>
<tr>
<td>Space Requirements</td>
<td>Relatively small</td>
<td>Relatively large</td>
</tr>
<tr>
<td>Database Design</td>
<td>Highly normalized with many tables</td>
<td>Typically de-normalized with fewer tables; use of star and snowflake schemas</td>
</tr>
</tbody>
</table>

Figure 1: OLTP vs OLAP Systems

Twitter
- User Privacy:
  - Ability to identify a user from a comment or image by third parties.
  - Option to display the geographical location where a comment or image was posted from.
  - Utilization of users’ personal information in order to associate certain advertisements with them.

- Improvements:
  - Parallelization using multithreading.
  - Design of a Graphical User Interface.
  - Crawler update to sequentially gather users using a file.
  - Crawler update to modify the tool’s configuration from within the application.
  - Crawler update to store incidents in a log file for later use (analysis or debugging).

Conclusions
- Use of a distributed cluster of machines to store and manage large amounts of data.
- Need for parallelized data collection due to the constantly increasing amounts of data that social networks process.
- Ability to connect to a social network using accounts from different networks.
- Ability to simultaneously collect user’s data from all the social networks in which they use the same account.
- Proactive critical infrastructure protection capability.
- Ability to enhance organizational monitoring systems to mitigate the insider threat.

Problems
- Time-consuming data management due to conventional relational databases.
- Delays in the data mining mechanisms due to lack of parallel processing.
- Need to upgrade existing mechanisms in order to make use of the latest API versions.

Table: OLTP vs. OLAP

Hadoop Ecosystem
- ETL Tools
- BI Reporting
- RDMS
- Pig (Data Flow)
- Hive (SQL)
- Sqoop
- MapReduce (Job Scheduling/Execution System)
- HDFS (Hadoop Distributed File System)

Figure 3: Hadoop ecosystem

Final Twitter Crawler
- User Privacy:
  - Ability to display user’s activity to third parties.
  - Ability to display video’s information (view count, likes, etc).
  - Connection with Google accounts.
  - Shared accounts with Facebook and Twitter.

- Improvements:
  - Updates and improvements on YouTube’s API responses.
  - Parallelization using multithreading.
  - Changes on the data stored in the data warehouse.

References