

Abstract

Because Absolute streams in such a wide diversity of formats and via a number of different entry points, rather than just web based, it is very awkward to simply build in a player-based heartbeat system as discussed in the section "Live Hours for Streaming – Flash" in the BBC document. We shall, therefore, be processing all our statistics from log files which are digested within 48 hours of broadcast. This is in correlation with the methodology discussed in section "Live Hours for Steaming – Real & WM" in the same document.

Policies

Each platform is processed separately using a plugin to our processing system, data extracted and inserted into a aggregated database. Reports are then done by running queries against this database.

Absolute reject any log records that meet any of the following criteria:

- Any log file entry generated as the result of a http HEAD verb
- Duration less than 6 seconds
- Bytes transferred = 0
- The source IP originates from Absolute's internal network
- The source IP originates from Absolute's external DMZ network (relaying another server)
- The source IP originates from University of Southampton's experimental relay station for IPv6
- Duration is greater than a day (since each log is rotated on a per day basis this is considered a corrupt entry)
- If the unique_id is a distribution-server player-id (uuid={000000...}) [applicable to WMS only]

These policies should bring us in congruency with the BBC policies suggested in the aforementioned document.

Reach

Absolute's strategy on reach remains the same as before. We analyse the WMS logs and determine percentage uniques to listeners hours and apply to other statistics where unique player identification is not possible. I've included the full details on this below for reference (excerpt from document "How we solve the hours/unique players for streaming" written by myself, 24/03/2009).

Ben Matthew, 30/06/2009

Unique Listener methodology (Currently unpublished)

Unique listener numbers reported by the system are considerably more complicated.

Neither shoutcast nor flash log any data related to unique listeners or "players".

Windows Media generates a "UUID" identifier for each instance of Windows Media Player installed anywhere in the world. We can use this identifier to spot particular people, on particular computers who tune in multiple times to the stream. However this is further complicated by the fact that Windows Media can be made to appear "anonymous"; this is where the reported UUID is set to something generic and does not, in fact, reflect a unique player.

Our methodology works like this:

1. Consider "Set A" which is the total Windows media listenership for a particular time period, "p".
2. Within period "p" there will be X players identifying themselves uniquely (either singly or multiple times) and Y players identifying themselves anonymously. [Total listeners = X+Y]
3. We can construct two subsets of "A", A1 and A2 where A1 = {set of identified players} and A2 = {set of anonymous players}
4. Within subset {A1} we can look for repeated instances of the same identifier, then group the total listener hours by those player ids. So instead of "600 hours spread over 1000 sessions" we can see "600 hours spent listening by 100 (unique) players". We take the unique listeners and refer to it as "U". However this is ONLY within subset A1.
5. For A1 we can sum the total listener hours, lets call that sum: A1H
6. For A2 we can sum the total listener hours, lets call that sum: A2H
7. The ratio $(A1H+A2H)/A1H$ gives us a ratio of IDENTIFIED players versus unidentified players, lets refer to that as "uR". Now we can expand the unique players, found in stage 4, over the entire set of "A" by multiplying U by uR. Lets refer to this number as uA.

8. We can now construct a ratio [within Set {A}] between total listenership hours and uniquely identified players. I.e.: $\text{Hours}\{A\} / uA$. Lets call this ratio "R"

9. We can now extrapolate this information across platforms. So if Shoutcast returns 1000 hours of listenership we can use ratio "R" to approximate what we'd expect the unique listenership count to be; i.e.: $\{\text{Shoutcast hours}\} * R =$ Unique shoutcast listeners. Same applies for flash.

Notes

If we consider Shoutcast to be Set B and Flash to be Set C then it should be noted that for UK-only data the ratios involved work by discounting non-UK tagged entries in the database BEFORE the subsets of A1 and A2 are constructed. Thus the ratio "R" is different dependant on geographic relationship.

If the total size of {A}, {B} or {C} is too small then the ratio "R" is considered unreliable and the analysis is stopped. Thus "p" must be significant in size (of at least 3 days currently) to make these figures work and, furthermore, dismisses the possibility of uniquely identifies statistics from the much, much smaller sets of data collected from the other streaming platforms.