

The General Anatomy of Log Files

Log Files

In general terms, a log file is a list of events that have taken place on an application, usually as a result of the interaction of a user with that system. Each line in the log file contains information pertaining to a specific event.

Each line of a log file must, as an absolute minimum, contain the following information:

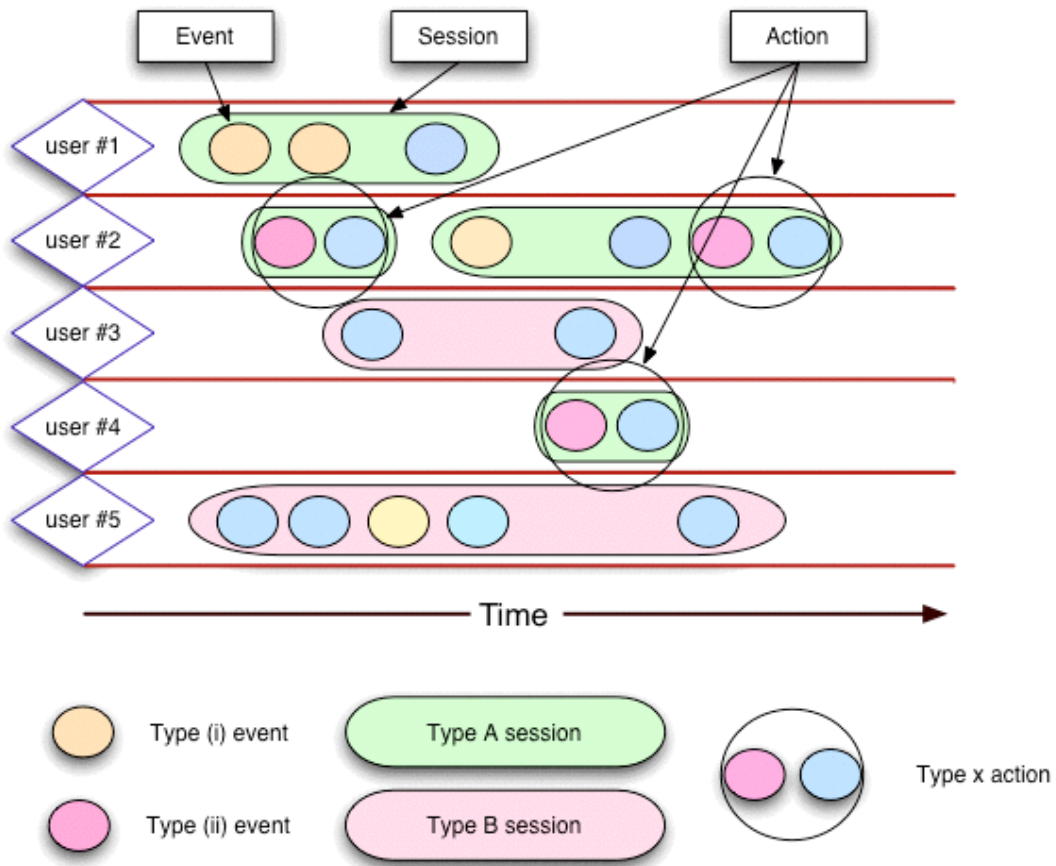
1. Time and date - usually the time of the occurrence of the particular event logged.
2. The event type - in a web log, for example, this is the html page served.
3. Some sort of event initiator identifier. This could be:
 - a. A session ID. or
 - b. A user ID, which can then be used to generate a proxy for a session ID, based on a set of assumptions about session timeout or similar.

A session is a sequence of events that can be attributed to a particular user over a particular time period. Sessions may themselves be typed, or categorised within the log file: for example, web sessions and wap sessions.

Users themselves can be categorised or typed within the log file, for example by geographical location, or status.

Finally, sequences of events within a session may constitute a whole action (described in commonsense terms, such as sending an email) initiated by the user.

What is a log file a picture of?



The diagram above illustrates a collection of users initiating events on an application over time. Each event:

- occurs at a particular time
- is of a certain type
- belongs to a session of a certain type, and of a certain duration
- is initiated by a user
- may or may not be part of an action-sequence

Understanding the relationship between the log file and the events it reports on is fundamental to understanding how to derive meaningful reporting information from that log file.

Reporting

From the diagram, we can see that there are several starting points for reporting.

For a given time period, and across that time period, we can produce reports showing:

1. User reporting
 - a. Number of users / active users (by type)
 - b. Number of sessions per user (by session type)
 - c. Total / mean duration of sessions per user (by session type)
 - d. Number of users (by type) initiating events (by type)
2. Session reporting
 - a. Total Number of sessions (by type)
 - b. Mean number of events (by type) per session (by type)
3. Event reporting
 - a. Total number of events (by type)
 - b. Number of events (by type) per session (by type)
 - c. Number of events (by type) per user (by type)
4. Action reporting
 - a. Total number of actions (by type)
 - b. Number of sessions (by type) including actions (by type)
 - c. Number of users (by type) initiating actions (by type)

Database Efficiency considerations

Application log files contain a lot of information, most of which is not going to be used for reporting. In building the database from which the reports will be generated, it is important, therefore, to include - as far as is possible - only data that will be used to generate reports. In general, we do this by aggregating the data over time.

First, all data is aggregated up to an hour-by-hour level: we throw away all information about the time an event took place except that it occurred with a particular one-hour timeslot. Total users and sessions are normally tracked hour-by-hour in this way.

Second, for individual users and event types, we aggregate all information up to the time period over which the report is run. Unless a watch is set up, there will be no data on individual users or events on an hour-by-hour basis.

A watch on a particular user or event sets up a database for that user that holds information down to an hour-by-hour level. For users, this might be as simple as a count, by hour, of the number of sessions or events of all types for that user. Or it might count specific types of events for that user, again at an hourly level.