Spam Filter Architecture

By Michal Young (adapted by Stuart Faulk)

Architecture Exercise

Context

The ABC Company produces a variety of internetworking utilities. In recent months, the number of ABC Company customers requesting help in eliminating spam (junk e-mail) has increased greatly, indicating a potentially lucrative market for a spam filtering utility. In initial discussions, a set of tentative product features and characteristics have been identified in internal brainstorming sessions and interviews with target customers. These tentative features and characteristics are not yet "requirements," pending an initial cost and feasibility study.

Your task

The next step for ABC Company is to make a rough cut at cost and feasibility for a product, or perhaps for multiple candidate products with different combinations of features and characteristics. This rough cut will depend not only on the requirements chosen, but also on the overall architecture of the product. Your task now is to produce an overall architecture that could form the basis of a cost and feasibility estimate now, and (after considerable refinement) a product later, if the project is approved.

Tentative requirements and characteristics

- ABC Company products are based on Internet standards. In particular, current ABC internet tools are based on the RFC 822 standard for mail addresses, the SMTP mail transport protocol for sending and forwarding mail, and the POP3 and IMAP standards for receiving email. ABC does not attempt to compete in the market for proprietary LAN-based email products.
- A successful spam filtering utility must be able to apply filtering criteria based on sender, recipients, and message content. Some filtering criteria may depend on information that is not entirely present in the message, but can be obtained using internet access. An example is that a filtering criterion may consider whether the sender's internet domain responds to "finger" requests.
- Spammers are constantly innovating to slip through current spam filters. It is therefore considered very unlikely that a sufficient set of spam filtering criteria can be identified once and for all. Rather, a spam filtering product must be updated on a very fast cycle, ranging from twice a year to twice a month.·
- Collaborative mail filtering has been raised as a possibility for establishing product differentiation. An example of collaborative mail filtering would be a filtering criterion that depends partly on the number of members of a group of users who receive an identical or near-identical message.
- Currently about 40% of ABC's customers use their products on or in conjunction with laptop computers, often with mail client programs such as Eudora, Netscape, or Outlook. Interoperability with these client programs is considered essential for product success.
- A small but rapidly growing portion of ABC's customers read their email on small hand-held devices such as PDAs or smart phones; this is regarded as an important market for future growth.
- About 30% of ABC's customers use corporate or organizational mail servers, while the remainder use accounts provided by internet service providers such as Worldcom or AT&T.

· The old joke about the hikers and the bear is applicable here. Two hikers observed a bear running toward them. One began to put on his tennis shoes. The other said, "Don't be silly, you can't outrun a bear." The first replied, "I don't have to outrun the bear, I only have to outrun you." It is worthwhile for a spam generator to defeat a new spam filtering criterion only if that filtering criterion is widely used. Therefore, a filtering product that initially has a small market share need not provide filtering criteria that cannot be defeated by the spam generators, it need only provide filtering that is somewhat stronger than that provided by the spam filters with larger market share.

Expected outcome

Please provide your answer in two parts:

Part 1: (due week 2) Part 1 of the Spam Filter Exercise focuses on understanding the problem in architectural terms. Please answer the following questions and be prepared to present and discuss your answers in class next week.

1. What is the business rationale that drives the overall design? If you believe that the business rationale is unclear or ambiguous, choose the one or ones you believe are appropriate and state why you choose them.
2. What are the architectural design goals? If you think there are several, pick the top two or three priorities. Briefly say why you believe your choices reflect the appropriate design goals (hint: given the business rationale you chose).
3. What set or sets of components and relations should you use to represent the architecture and why? I.e., tell me which sets of components and relations you would use to represent the Spam Filter Architecture and why you have chosen a particular set (for each set chosen if there is more than one). Note that your rationale should be in terms of which architectural issues you are designing for and which respective decisions you want to represent and communicate.
4. For each architectural "view" you would create, tell me how you would verify the "goodness" of a design relative to your design goals. I.e., what questions would an independent evaluation team ask about the design to determine if you had met your design goals. Please be as specific as possible in terms of what inputs the evaluators would use (i.e., what document content), what properties they would check for, and how they would determine whether those properties were satisfied.

Part 2: (due week 3) You should produce an architectural "sketch" of candidate spam filtering tool architecture, possibly with variants. Your sketch should identify major components and interfaces (connections). Your sketch should include

- A brief description of the functionality of each major component. This should be sufficiently clear that it can answer questions of the form "is functionality XXX present, and if so, where is it?" Together with the interface information, it should also be sufficient as a basis for a very rough estimate of the development cost of each major component.
- A brief description of each interface between major components.
- Optionally, a description of alternatives that you considered and rejected, with your rationale for doing so.
- A brief description of the overall organizing principles governing your proposed architecture.
- Your rationale for the overall design – this only needs to be a brief overview with pointer to your answers from Part 1 of the exercise.

It is expected that this can be done in about 5 pages.

Which architecture?

Your textbook and some of the readings will stress that there are several possible architectural views of a system: run-time process organization, module decomposition and dependence, and others. So, which architecture(s) are we asking you to describe? We aren't saying. You must choose an appropriate architectural view or views, considering the purpose to which your architectural sketch will be put.

Don't Panic

You may feel uncomfortable doing this exercise at this point in the course. Don't panic; do your best. We know that many of the relevant topics of the course have not been covered yet. Wrestling with this problem now will, we hope, prepare you to get more out of our readings and discussions of those topics.