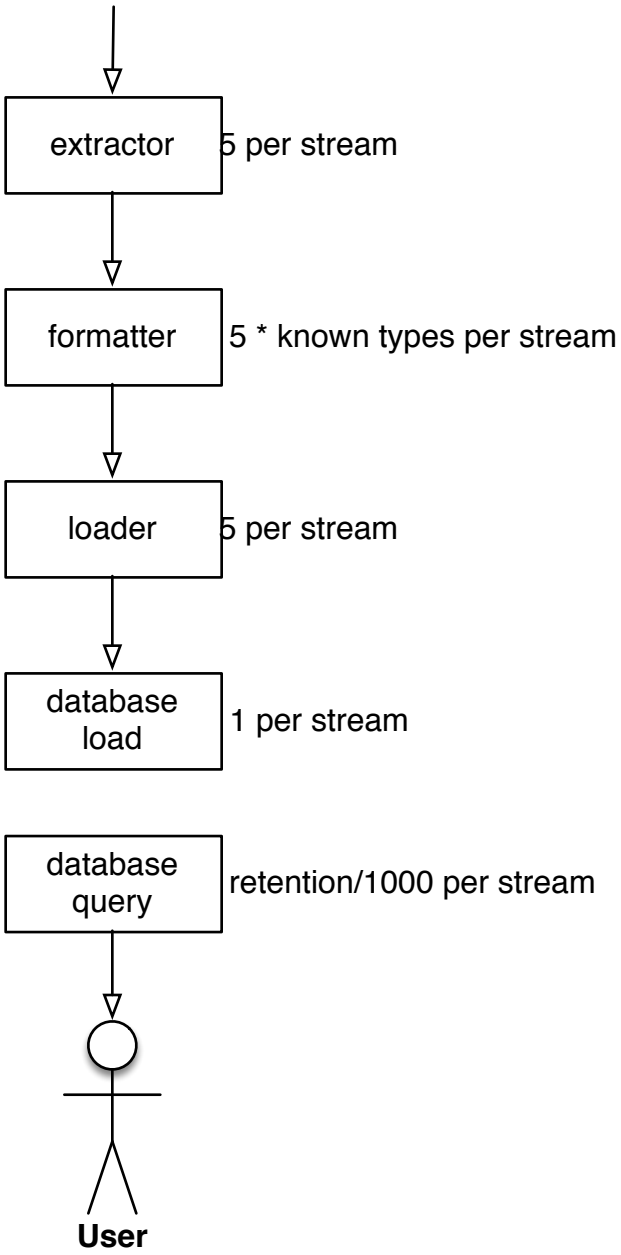
For the following problem we assume that network cost between steps can be ignored. Costs for steps are given in arbitrary compute units.

A log mining system has the following structure:



The extractor step connects to a log stream and captures the log records in the source's preferred export format once per second. Each stream costs 5 compute units to capture.

The formatter step normalizes the format of the captured records and generates some number of load records from the log record. While the increase in data is negligible, each stream type supported by the system will increase the cost of

processing all streams. Each stream costs 5 * <number of supported types> compute units.

The loader step converts the normalized load records into a binary format that enables the data to be loaded faster than insert statement. Each stream costs 5 compute units.

The database step on the load side adds the binary data to the database. Database queries happen at a different rate than log record extraction and have a cost based on the size of the database. Only one query will ever execute at a time. Each stream cost 1 compute unit.

The initial deployment is scaled to support 1 stream type from 10 sources on a single host. Data retention is 1000 seconds, resulting in a data volume of 10000, and therefore a query cost of 10 units. This deployment can be done using only a single host since each host provides 200 units of computation.

Extractor cost 50 (5 units * 10 streams)
Formatter cost 50 (5 units * 1 type * 10 streams)
Loader cost 50 (5 units * 10 streams)
Database load cost 10 (1 units * 10 streams)
Database query cost 10 (1000 sec * 10 streams / 1000 factor)

The database load and query steps must be done using a single host (the case when sharding and multimaster can't be used), what is roughly largest number of streams a deployment could support while retaining the 1000 second data retention?

What if the retention requirement was increased to 3000 seconds?

Assuming the 1000 second data retention. If there were 20 streams to collect, containing 2 types of information, how many 200 unit hosts would you suggest be used to process the data? How might you map the processes onto these physical hosts? Why did you choose this mapping and host count?

You do not need to use the least number of hosts. Splitting the output from a step to multiple hosts does not cost compute but please label the number of streams traversing each edge in your diagram.