# Projects

This is an un-ordered compendium of projects and project ideas.



### Non-java Clients

We want to improve the client libraries for the major languages (ruby, python, c++, etc). Some of these don't yet have a 0.8 compatible library available and for others the client is somewhat limited and could be improved.

### Audit Trail

LinkedIn has an "audit" application that checks the correctness of the data pipeline by comparing published and consumed messages. It would be nice to get this open sourced as well as make a number of improvements to it.

### Performance

There are a number of projects that fall under the general bucket of performance improvements that aren't called out elsewhere:

- Improved I/O management - Move the flush out of the main thread and avoid linux file locking.
- Mmap log for writes to improve small write performance.
- General data-driven profiling and perf improvements
- Memory hardening. Now due to async requests it is possible to OOM the server. It would be good to write some torture tests for the server and work on hardening its memory usage patterns.

### Cleanup and Refactoring

- Purgatory rewrite - The current data structure that handles async requests is a bit hacky and could be improved
- Kafka API split - Currently we maintain a single class (KafkaApis.scala) that has all request handling logic. As we add apis this is unsustainable, we should have one handler class per API just to help shrink and separate this giant lump of code.
- Move build to maven

### Exactly once producer semantics

Now that we have replication it would be possible to implement exactly-once producer semantics.

### Log Slurper

For people who want to publish Kafka feeds for existing applications that produce log files it might be nice to have something more sophisticated than the console-producer. This would be a process that ran in the background and tailed log directories and read and published formatted messages.

### Cluster Admin UI

It would be nice to have a simple web app that showed the state of the cluster--which brokers are up, what topics and partitions they replicate and lead, how much data they have etc.