A few questions to consider

These are a set of common questions which arise when designing systems in a distributed environment. This is not a complete list by any means, but they may be a useful set of *prompts* for issues to consider.

1. Fault tolerance

- What happens when a dependency starts failing? What if it begins failing *slowly*?
- How can the system *degrade* in a graceful manner?
- How does the system react to *overload*? Is it "well conditioned?"
- What's the worst-case scenario for total failure?
- How quickly can the system recover?
- Is delayable work delayed?
- Is the system as *simple* as possible?
- How can the system *shed load*?
- Which failures can be mitigated, and how?
- Which operations may be *retried*? Are they?

2. Scalability

- How does the system *grow*? What is the chief metric with which the system scales?
- How does the system scale to multiple datacenters?
- How does demand *vary*? How do you ensure the system is always able to handle peak loads?
- How much query processing is done? Is it possible to shape data into queries?
- Is the system replicated?

3. Operability

- How can features be turned on or off?
- How do you monitor the system? How do you detect anomalies?
- Does the system have operational needs specific to the application?
- How do you deploy the system? How do you deploy in an emergency?
- What are the capacity needs? How does the system grow?
- How do you configure the system? How do you configure the system *quickly*?
- Does the system behave in a *predictable* manner? Where are there nonlinearities in load or failure responses?

4. Efficiency

- Is it possible to precompute data?
- Are you doing as little work as possible?
- Is the program as concurrent as possible? ("Concurrent programs wait faster.")
- Does the system make use of work batching?
- Have you profiled the system? Is it possible to profile in situ?
- Are there opportunities for parallelization?
- Can you load test the system? How do you catch performance regressions?