How to Select Technologies

Dec 8, 2015

The Problem

We've seen a lot of different systems:

- Storage: GFS, BigTable, Dynamo, databases, …
- Resource sharing: Mesos, Borg, EC2, ...
- Analytics: MapReduce, Spark, Dremel, Naiad, ...
- Serving: Tao, Unicorn, Druid, ...

How to decide which one to use when?

Key Considerations

1. Project architecture
2. Available alternatives
3. Workload envelope
4. Ease of management

1. Project Architecture

Software has to meet many conflicting goals

- Robustness
- Performance
- Security
- Time to market

Is there any way to improve in all of them?

YES: keep the software small.

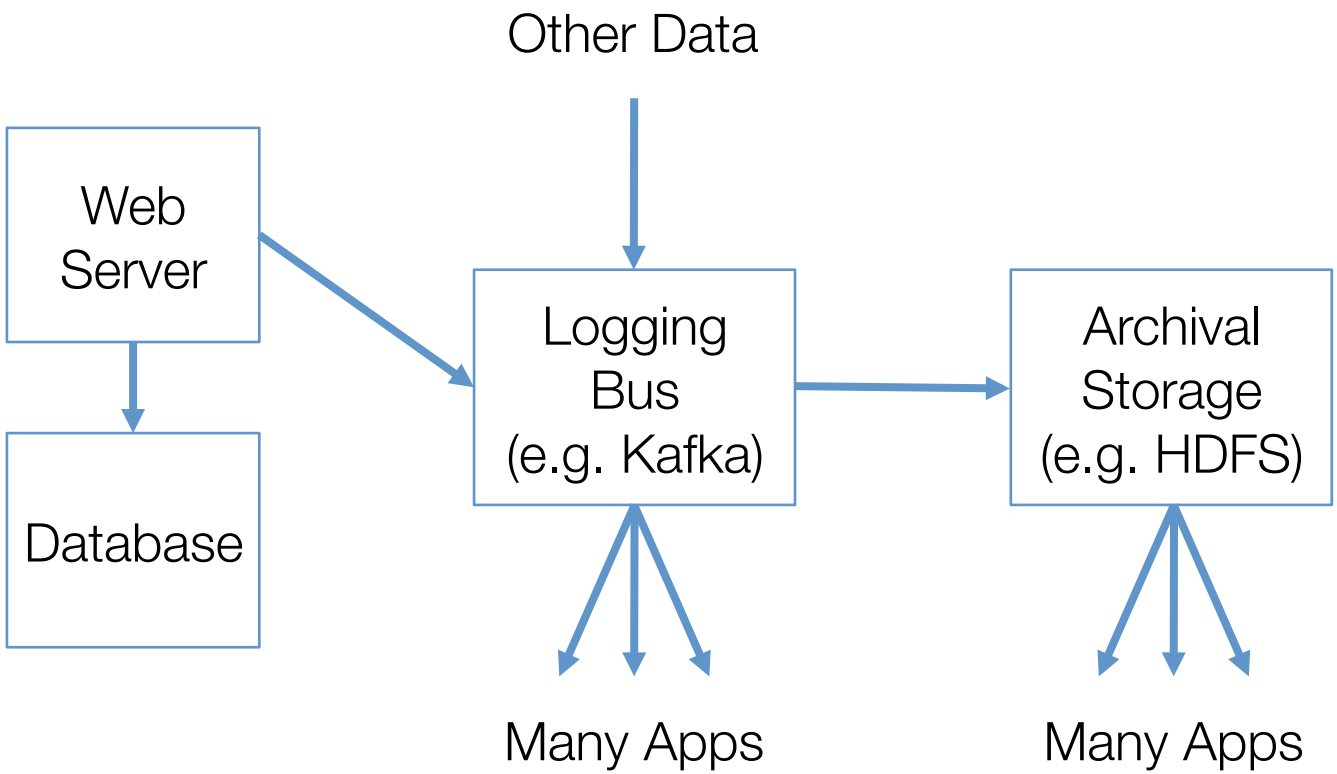
More Generally

Good design needs to be easy to understand and easy to change

Key considerations for big data systems:

- Is the data easy to access from other tools?
  - E.g. HDFS, S3, Kafka
- Do apps need to know details of the system?
  - Consistency, availability, performance quirks
- Is it easy to plug in existing code?

Typical Designs



2. Available Alternatives

Find commonly used systems for a problem

- May not obviously be in the same category, e.g. database vs key-value store

Requires quickly parsing “jargon” for each one

- Similar to reading the papers we covered

3. Workload Envelope

What workload characteristics does the system support?

- Scale in various dimensions (total data size, number of fields, length of records, etc)
- Performance metrics (throughput, latency, etc)
- Deployment environment (e.g. multi-datacenter)
- Consistency and availability

How to Find Workload Envelope

Design an evaluation workload you can test on different systems

- E.g. load and query fake data

Read about other deployed use cases

4. Ease of Management

Many factors matter for real-world use:

- How easy is the system to install or upgrade?
- Can you find people trained in using/managing it?
- How can you monitor it / tell it's working?
- How can you troubleshoot if something's wrong?

Example

Summary

- Define the problem you want to solve
- List most common alternatives
- Design and run an evaluation task (for both workload and management actions)
- Write down and share the results