Umeng Analytical Arch

Zhang Yan 2014.02

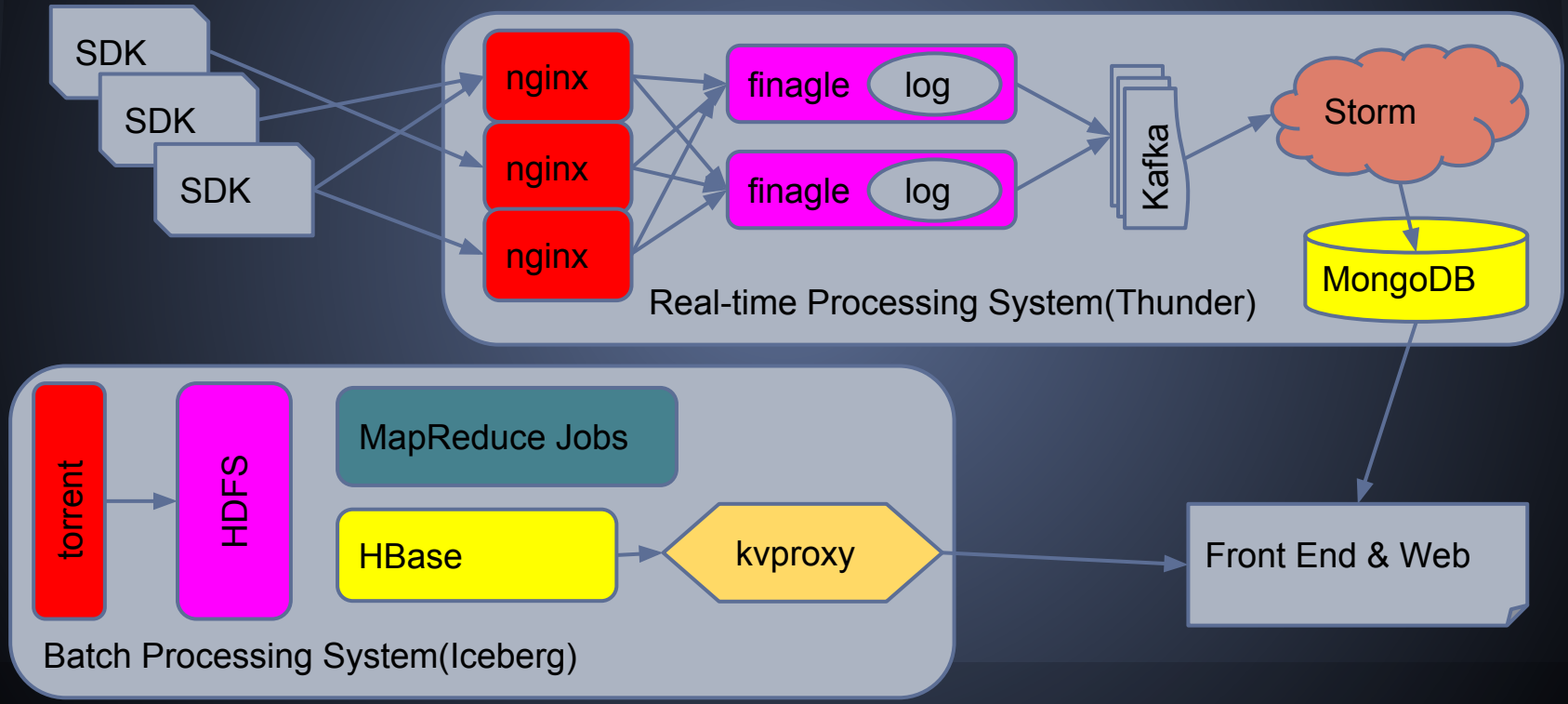
Outlines

- what are we doing
- how are we doing
- real-time system
- batch system
- lambda arch
- future & challenge

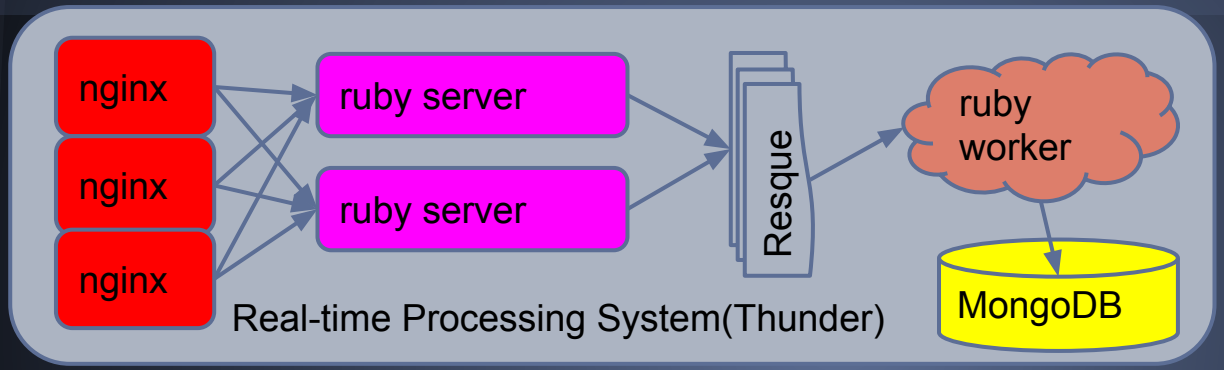
What are we doing

- founded on 2010.4
- focus on mobile app analytics (and share)
- 100k apps, 600 million devices
- 1PB data, 1TB increment per day
- 1 billion messages per day
- real-time: 15 nodes, 60k QPS (data skew)
- batch: 100 nodes, 500+ jobs

How are we doing

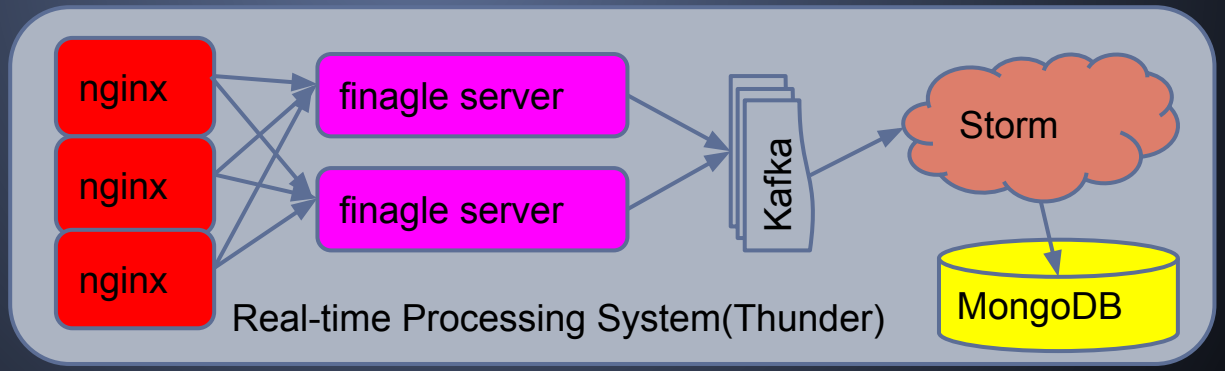


Real-time system



Evolve ...

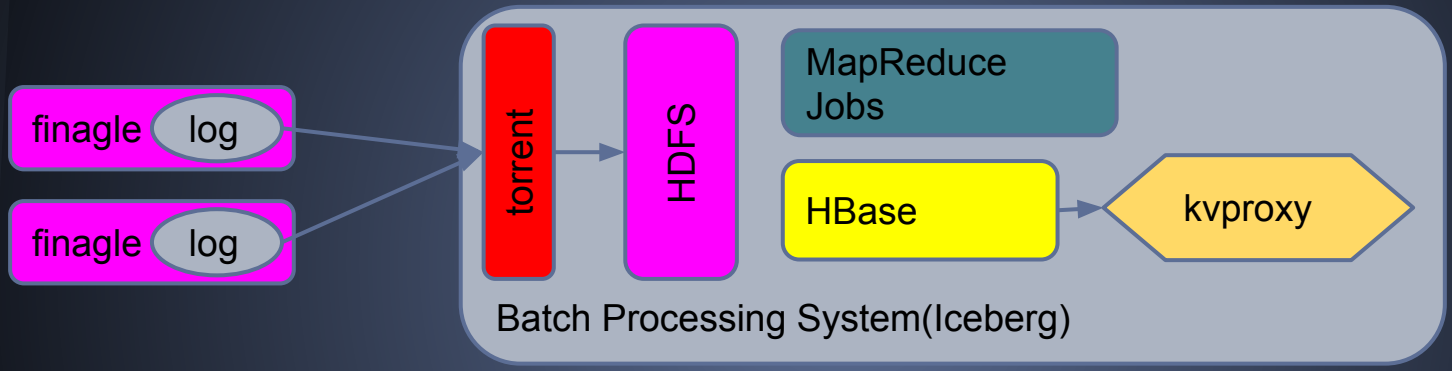
To ....



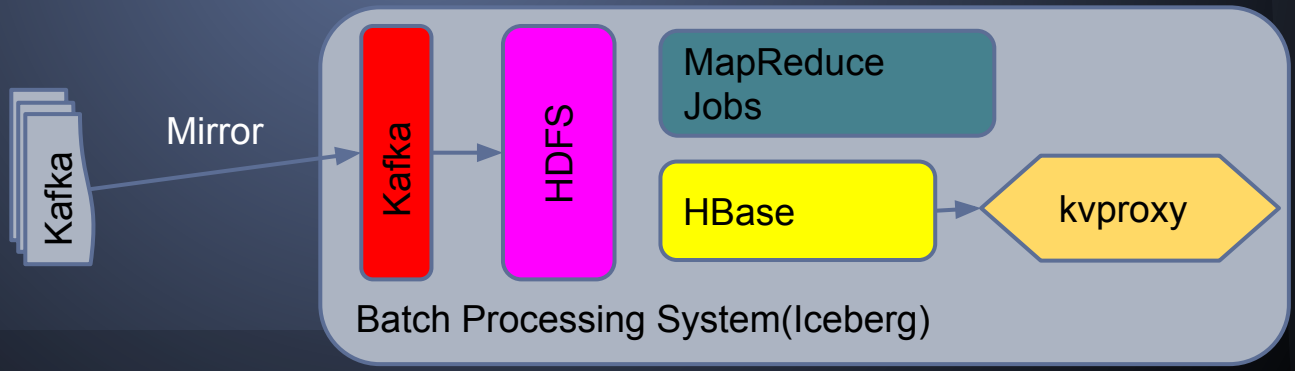
Real-time system(cont.)

- technology we use
  - java, scala, ruby
  - nginx, resque, mongodb
  - finagle, kafka, storm
- challenge we face
  - data skew
  - mongodb
    - global lock is bad.
    - table lock, but still good not enough.

Batch system



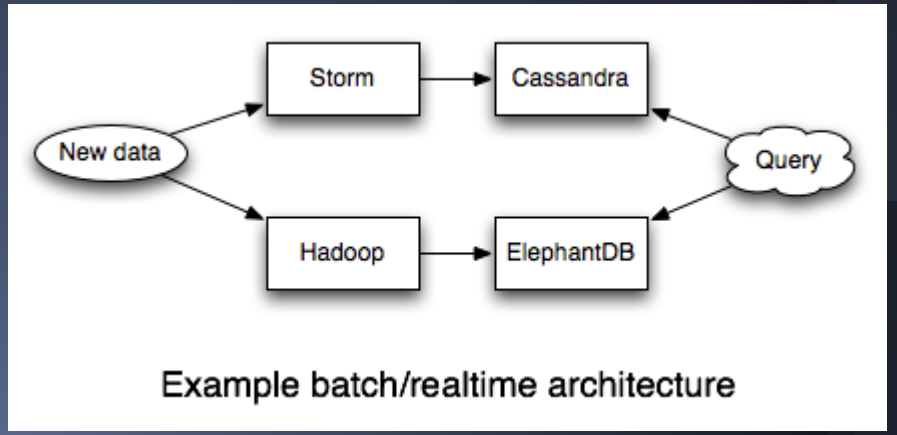
But we wish it could be ...



Batch system(cont.)

- technology we use
  - java, scala, python, c/c++
  - hadoop, hdfs, hbase, pig, hive, opentsdb
- challenge we face
  - data processing in a uniform way
  - optimize hbase usage(hash prefixed key, bulk load)
  - big amount of precomputation(HyperLogLog)
  - debugging, deployment, measurement etc.

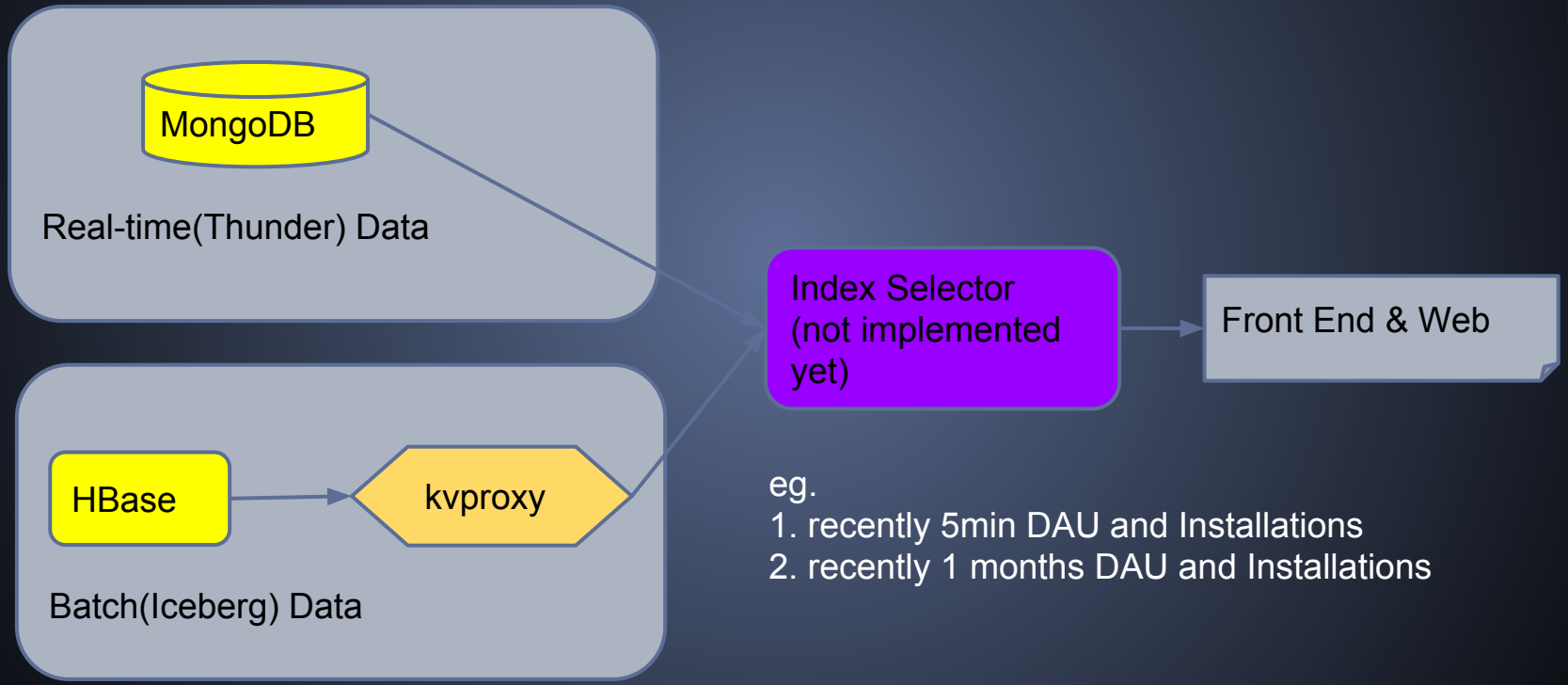
Lambda arch



Example batch/realtime architecture

- merge real-time and batch
- complete each other
- real-time
  - fast, imprecise, narrow scope
- batch
  - slow, accurate, wide scope

Lambda Arch(cont.)



eg.
1. recently 5min DAU and Installations
2. recently 1 months DAU and Installations

Future & Challenge

- just-in-time computation
  - druid, impala, spark etc.
- unify real-time & batch processing in one piece of code.
- data availability(internal & external)