CLOUD COMPUTING APPLICATIONS

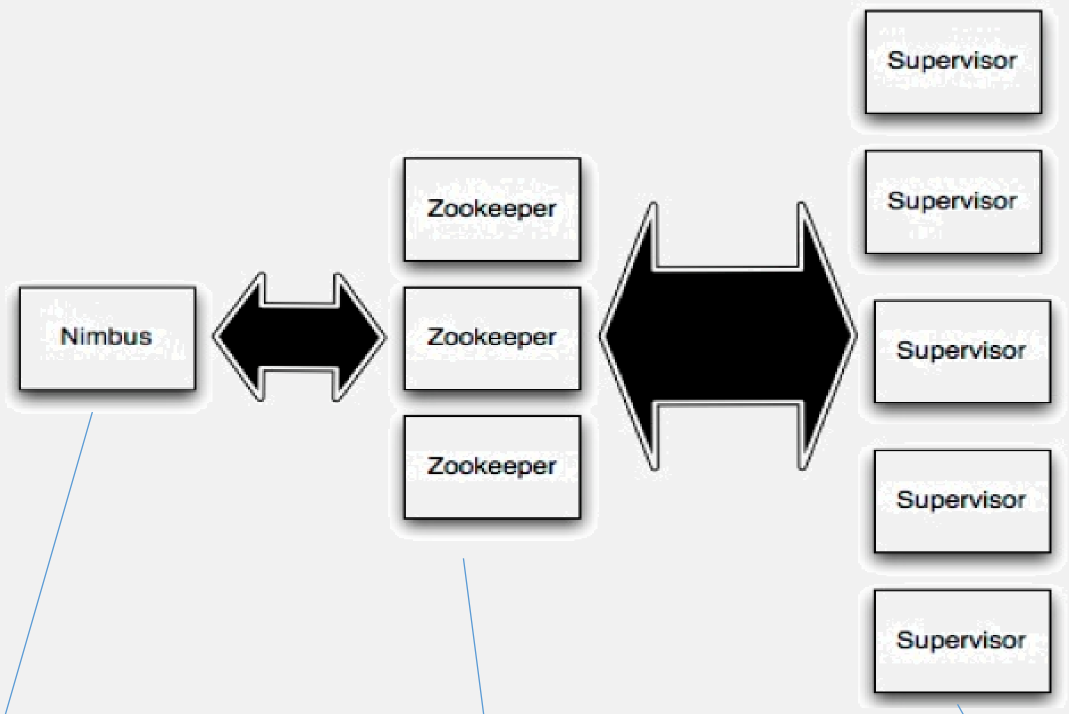
Storm Introduction: Bolts & Spouts

Roy Campbell & Reza Farivar

Apache Storm

- Guaranteed data processing
- Horizontal scalability
- Fault tolerance
- No intermediate message brokers
- Higher-level abstraction than message passing
- “Just works”
  - Hadoop of real-time streaming jobs
- Built by Backtype, then by Twitter, and eventually Apache open source

Storm



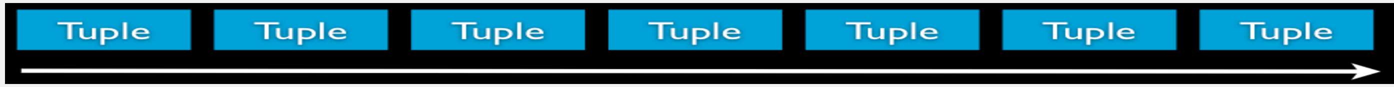
Master Node

Cluster Coordination

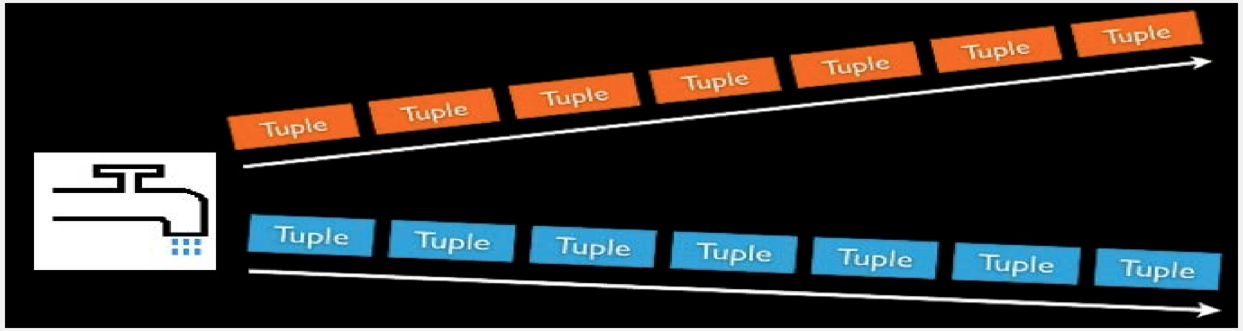
Worker Processes

Storm Concepts

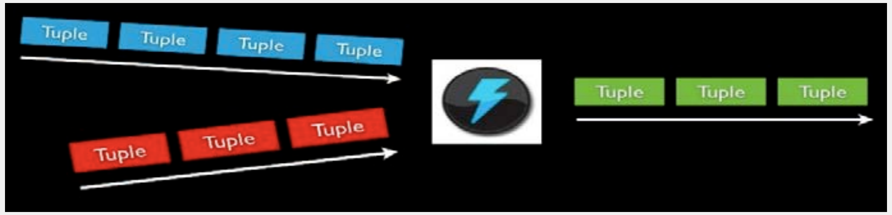
- Streams
  - Unbounded sequences of tuples



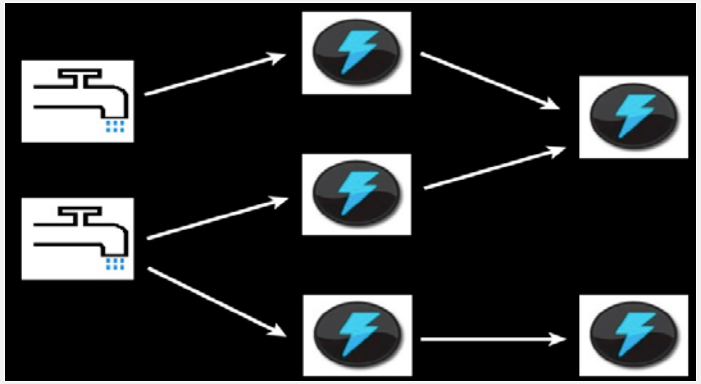
- Spout
  - Source of Streams
  - E.g., Read from Twitter streaming API



- Bolts
  - Processes input streams and produces new streams
  - E.g., Functions, Filters, Aggregation, Joins

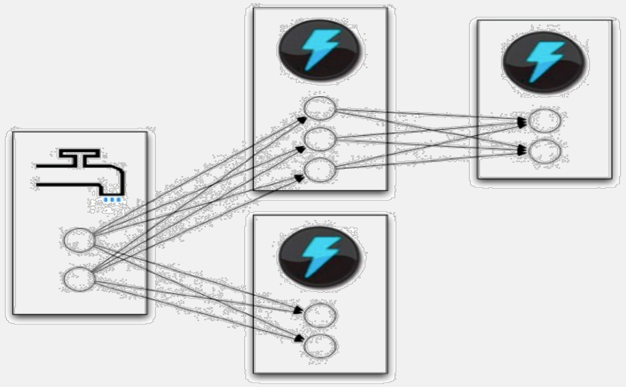


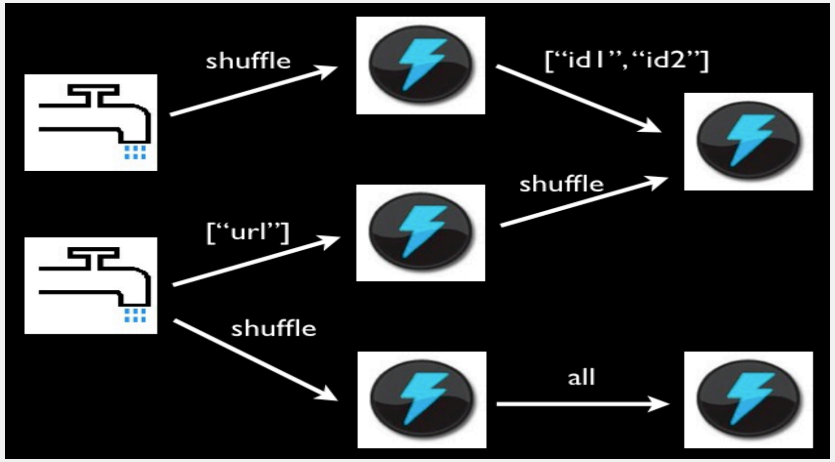
- Topologies
  - Network of spouts and bolts



Storm Tasks

- Spouts and bolts execute as many tasks across the cluster
- When a tuple is emitted, which task does it go to? → User programmable:
  - Shuffle grouping: pick a random task
  - Fields grouping: consistent hashing on a subset of tuple fields
  - All grouping: send to all tasks
  - Global grouping: pick task with lowest id




shuffle
[“id1”,“id2”]
shuffle
[“url”]
shuffle
all