15-826: Multimedia Databases and Data Mining

Extra: intro to hadoop

C. Faloutsos

Resources

- Software
- Map/reduce paper [Dean & Ghemawat]
- Tutorial: see part1, foils #9-20
  - [videolectures.net/kdd2010_papadimitriou_sun_yan_lsdm/](http://videolectures.net/kdd2010_papadimitriou_sun_yan_lsdm/)

Motivation: Scalability

- Yahoo: 5Pb of data [Fayyad, KDD’07]
- Problem: machine failures, on a daily basis
- How to parallelize data mining tasks, then?
2' intro to hadoop

- master-slave architecture; n-way replication (default n=3)
- ‘group by’ of SQL (in parallel, fault-tolerant way)
- e.g., find histogram of word frequency
  - compute local histograms
  - then merge into global histogram

```sql
select course-id, count(*)
from ENROLLMENT
group by course-id
```

By default: 3-way replication; 
Late/dead machines: ignored, transparently (!)
More details:

• (thanks to U Kang for the animations)
Conclusions

• Convenient mechanism to process Tera- and Peta-bytes of data, on >hundreds of machines
• User provides map() and reduce() functions
• Hadoop handles the rest (eg., machine failures etc)