Hadepot: A Repository of Big-Data Applications

Magdalena Balazinska

UNIVERSITY OF WASHINGTON

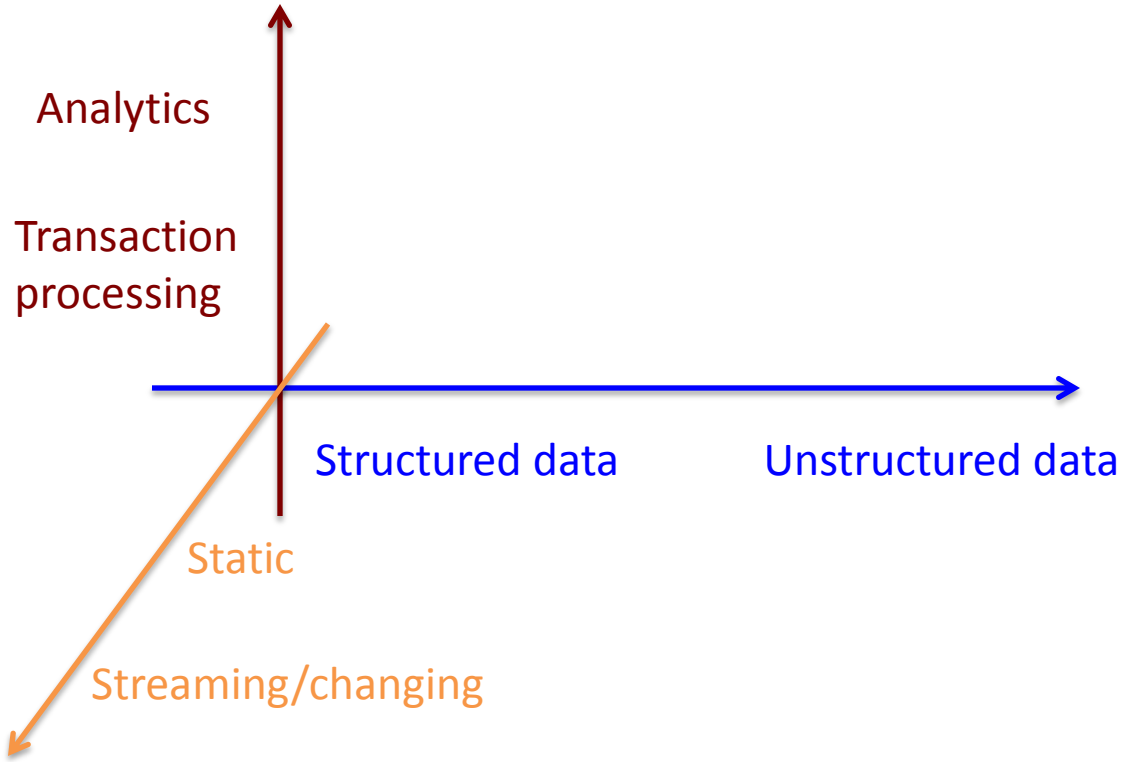
http://www.cs.washington.edu/people/faculty/magda

Benchmarking Using Real Applications

- Benefits
  - Ensures systems are tuned for real applications
  - As applications evolve, so do the benchmarks
  - Benchmarks can correspond to “grand challenges”
    - E.g, Discover new stars/planets
- Challenges
  - Incentivize contributions of applications
  - Maintain these contributions
  - *Categorize applications (rest of this talk)*

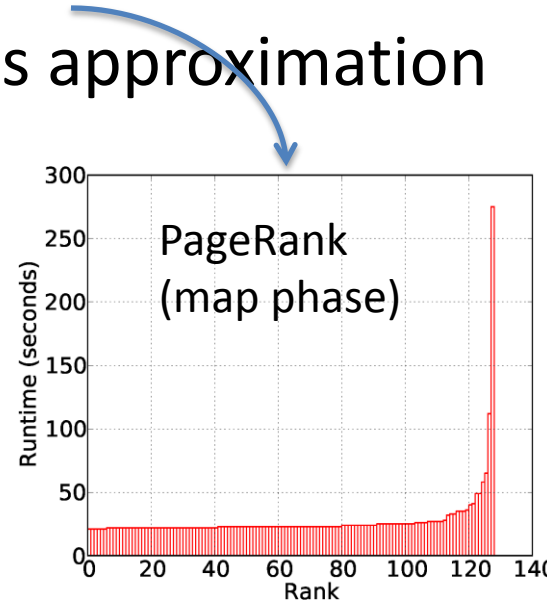
Creating a High-Level Taxonomy

- Cloud applications are diverse
- It is easy to create high-level categories



Discovering and Cataloging More Detailed Properties

- But apps have many additional properties
- For example, consider a big-data analytics app:
  - Complex workflow vs simple sequence of operations
  - Embarassingly parallel or tightly coupled (how?)
  - Easy to load balance or prone to skew
  - Requires precise answers or tolerates approximation
  - CPU/IO/memory intensive
  - Etc.



Hadepot Approach

http://nuage.cs.washington.edu/repository.php

- Repository of Hadoop applications
  - Accept any app with best-effort description
  - Let researchers play with the apps
  - Ask them for interesting properties they found
- Does this approach work: NO!
  - Lack of incentives to contribute
  - Approach is too labor intensive
- Proposed approach:
  - Develop an extensible and automated classification tool