Map-Reduce

Refinements
Implementations

Mining of Massive Datasets

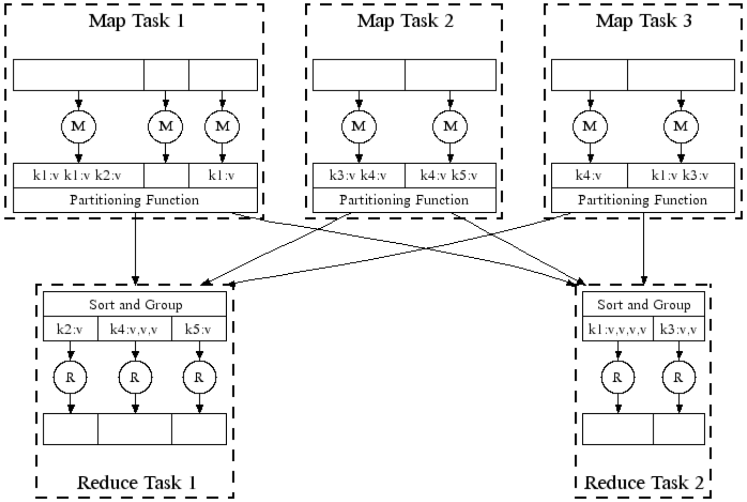
Leskovec, Rajaraman, and Ullman

Stanford University

Refinement: Combiners (1)

- Often a Map task will produce many pairs of the form (k,v_1), (k,v_2), … for the same key *k*
  - E.g., popular words in the word count example

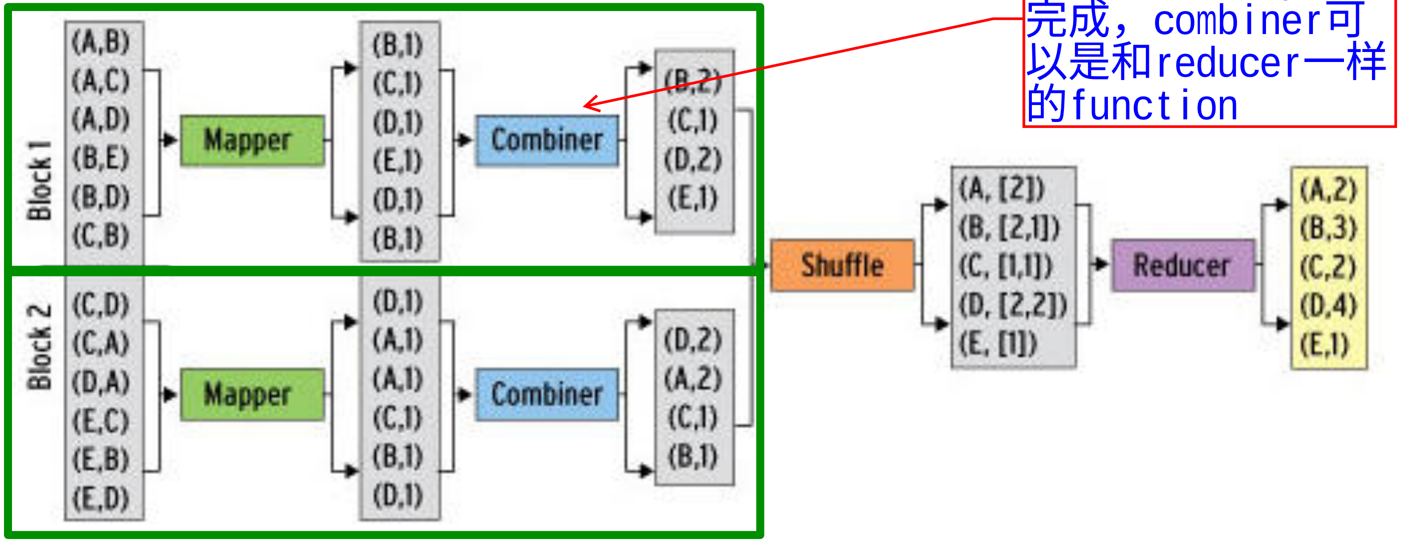
- **Can save network time by pre-aggregating values in the mapper:**
  - combine(k, list(v_1)) → v_2
  - Combiner is usually same as the reduce function



减少traffic，一些工作可以在map里面完成

Refinement: Combiners (2)

- **Back to our word counting example:**
  - Combiner combines the values of all keys of a single mapper (single node):



减少traffic，一些工作可以在map里面完成，combiner可以是和reducer一样的function

  - Much less data needs to be copied and shuffled!

Refinement: Combiners (3)

- Combiner trick works only if reduce function is commutative and associative

满足交换律和结合律

- Sum

- Average

这个可以记录count和sum，修改一下即可满足交换律和结合律

- Median

这个不行，中位数

Refinement: Partition Function

- **Want to control how keys get partitioned**
  - The set of keys that go to a single reduce worker

- **System uses a default partition function:**
  - **hash(key) mod *R***

- **Sometimes useful to override the hash function:**
  - E.g., **hash(hostname(URL)) mod *R*** ensures URLs from a host end up in the same output file

Implementations

- Google MapReduce
  - Uses Google File System (GFS) for stable storage
  - Not available outside Google

- Hadoop
  - Open-source implementation in Java
  - Uses HDFS for stable storage
  - Download: http://lucene.apache.org/hadoop/

- Hive, Pig
  - Provide SQL-like abstractions on top of Hadoop Map-Reduce layer

Cloud Computing

- Ability to rent computing by the hour
  - Additional services e.g., persistent storage

- E.g., Amazon's "Elastic Compute Cloud" (EC2)
  - S3 (stable storage)
  - Elastic Map Reduce (EMR)

Pointers and Further Reading

Reading

- Jeffrey Dean and Sanjay Ghemawat: MapReduce: Simplified Data Processing on Large Clusters
  - http://labs.google.com/papers/mapreduce.html

- Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung: The Google File System
  - http://labs.google.com/papers/gfs.html

Resources

- Hadoop Wiki
  - Introduction
    - http://wiki.apache.org/lucene-hadoop/
  - Getting Started
    - http://wiki.apache.org/lucene-hadoop/GettingStartedWithHadoop
  - Map/Reduce Overview
    - http://wiki.apache.org/lucene-hadoop/HadoopMapReduce
    - http://wiki.apache.org/lucene-hadoop/HadoopMapRedClasses
  - Eclipse Environment
    - http://wiki.apache.org/lucene-hadoop/EclipseEnvironment
- Javadoc
  - http://lucene.apache.org/hadoop/docs/api/

Resources

- Releases from Apache download mirrors
  - http://www.apache.org/dyn/closer.cgi/lucene/hadoop/
- Nightly builds of source
  - http://people.apache.org/dist/lucene/hadoop/nightly/
- Source code from subversion
  - http://lucene.apache.org/hadoop/version_control.html