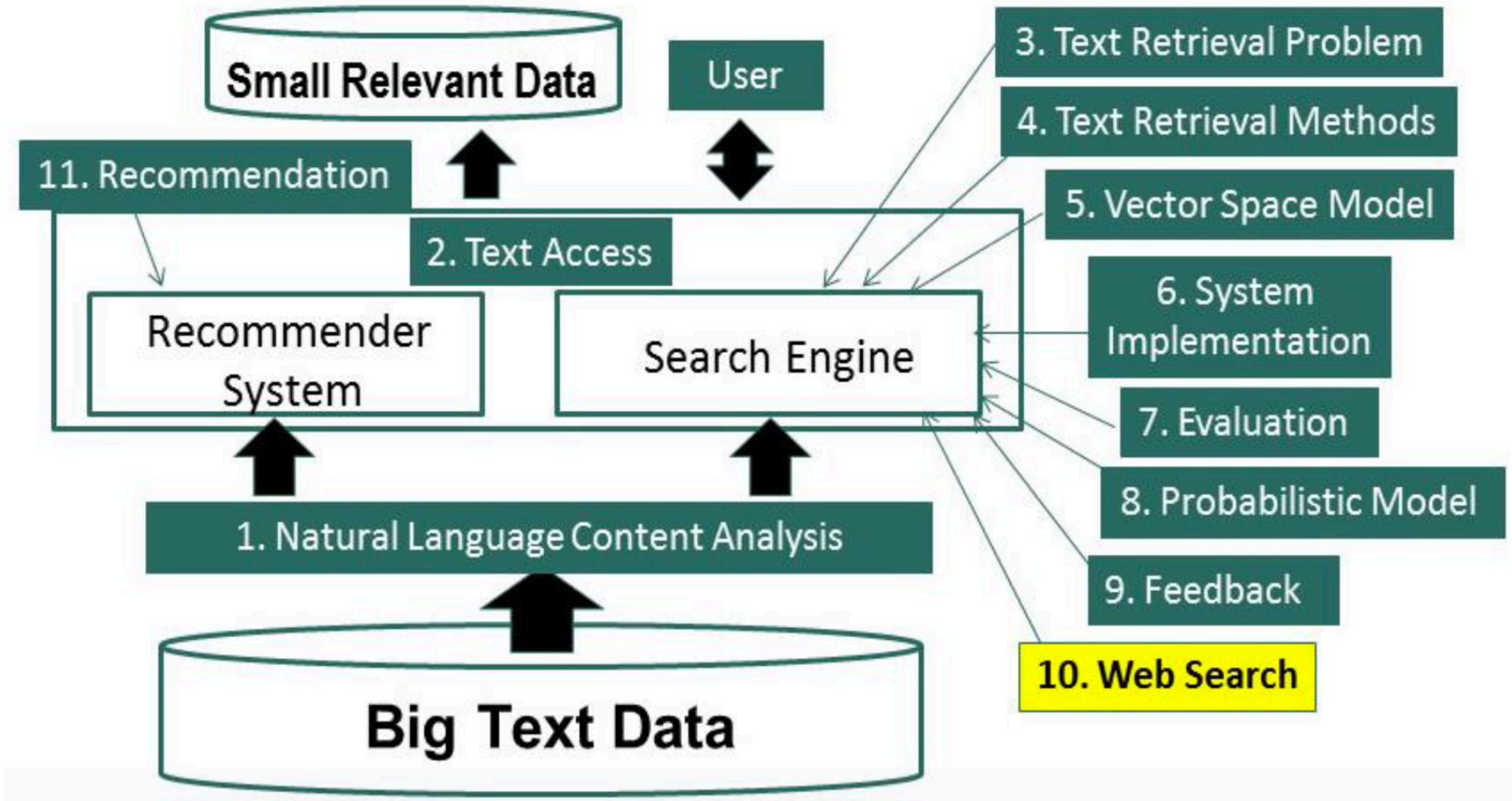
Information Retrieval & Text Mining

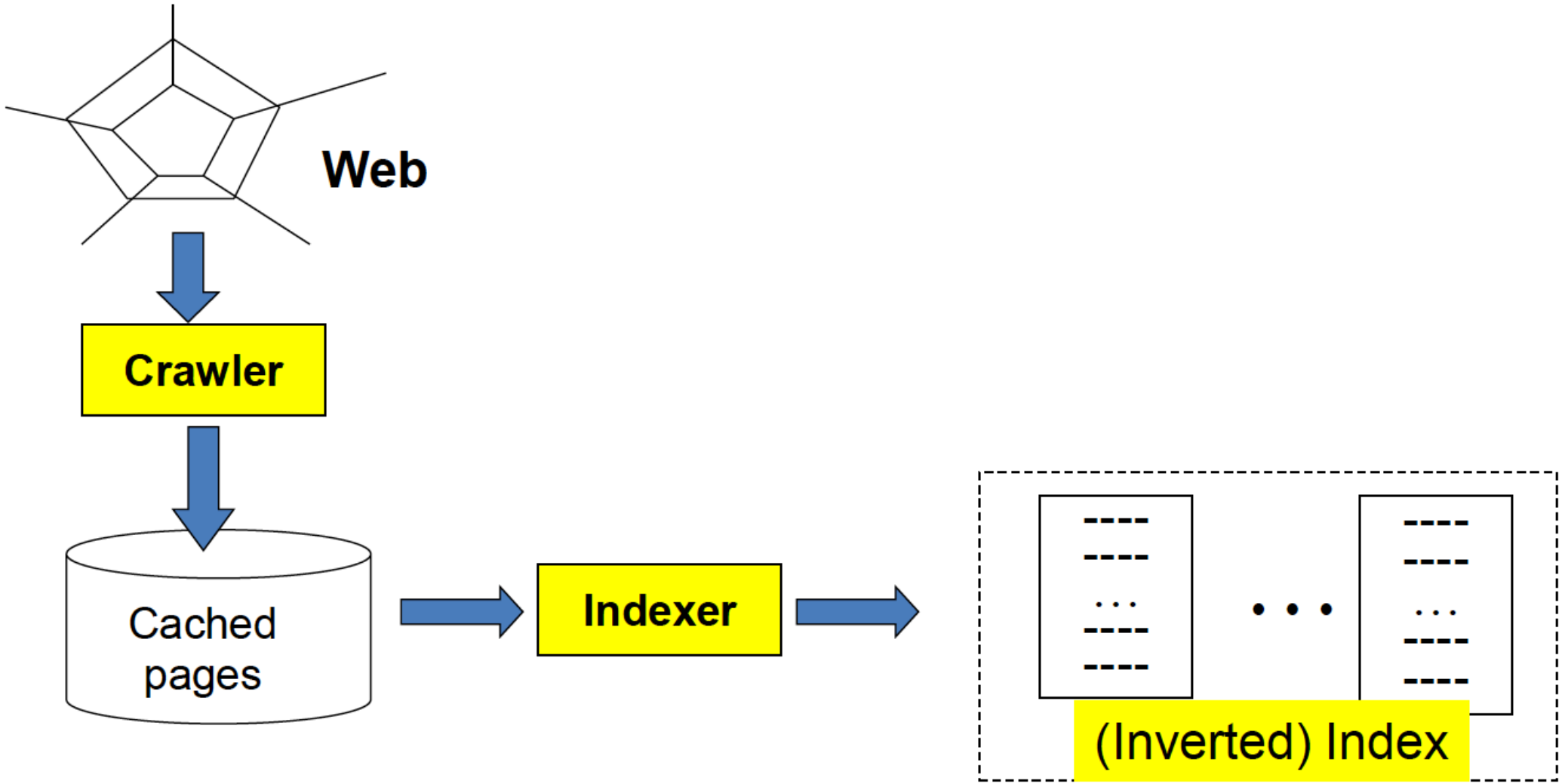
Web Search
Web Index

Dr. Saeed Ul Hassan
Information Technology University

Course Schedule



Basic Search Engine Technologies



Overview of Web Indexing

- Standard IR techniques are the basis, but insufficient
  - Scalability
  - Efficiency
- Google's contributions:
  - Google File System (GFS): distributed file system
  - MapReduce: Software framework for parallel computation
  - Hadoop: Open source implementation of MapReduce

GFS Architecture

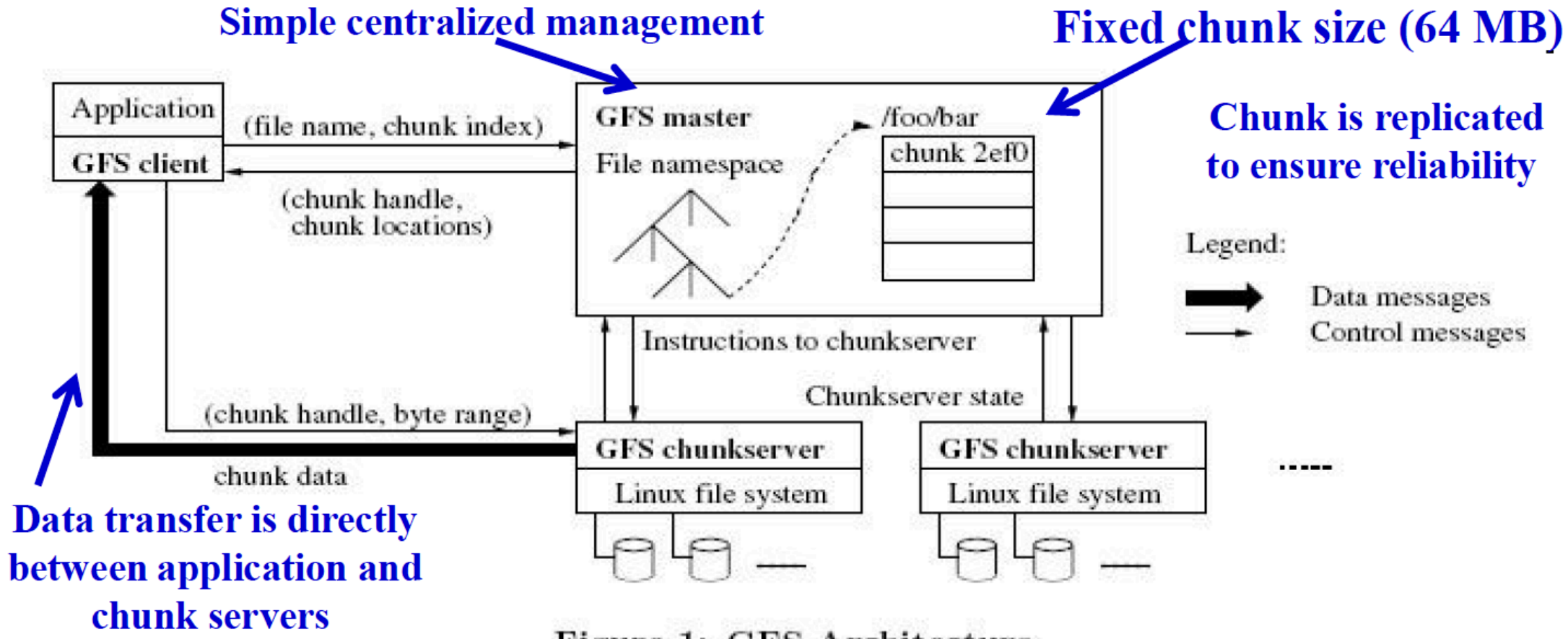


Figure 1: GFS Architecture

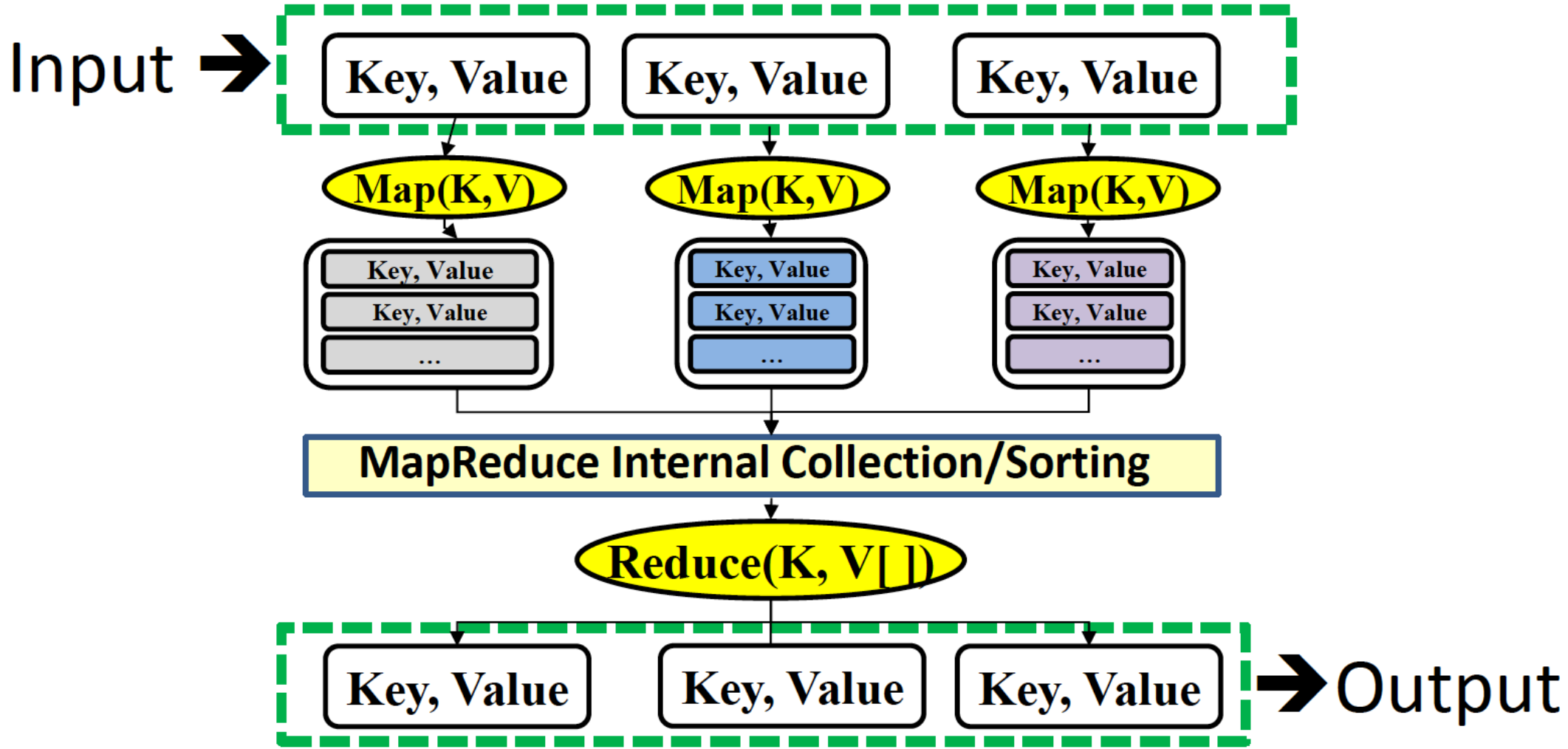
GHEMAWAT, S., GOBIOFF, H., AND LEUNG, S.-T. The google file system. In SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles (New York, NY, USA, 2003), ACM, pp. 29–43.

http://static.googleusercontent.com/media/research.google.com/en/us/archive/gfs-sosp2003.pdf

MapReduce: A Framework for Parallel Programming

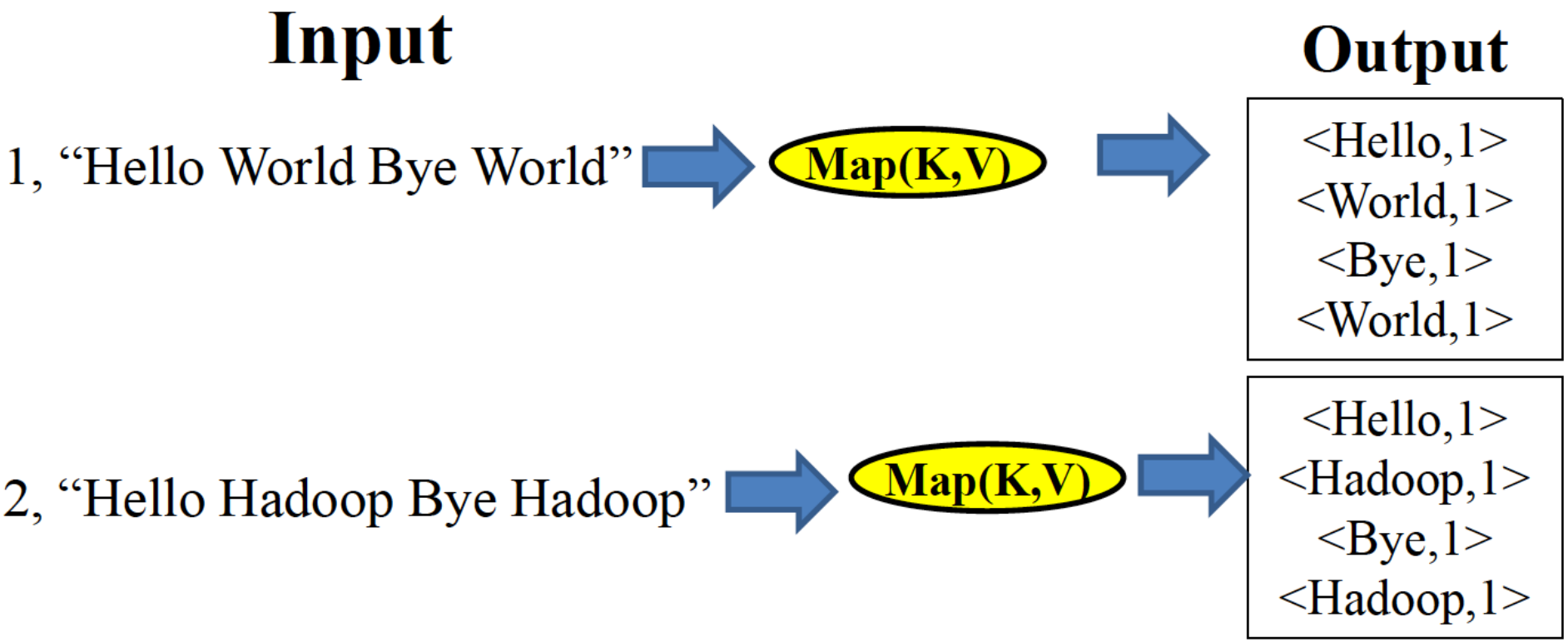
- Minimize effort of programmer for simple parallel processing tasks
- Features
  - Hide many low-level details (network, storage)
  - Built-in fault tolerance
  - Automatic load balancing

MapReduce: Computation Pipeline



Slide adapted from Alexander Behm & Ajey Shah's presentation (http://www.slideshare.net/gothicane/behm-shah-pagerank)

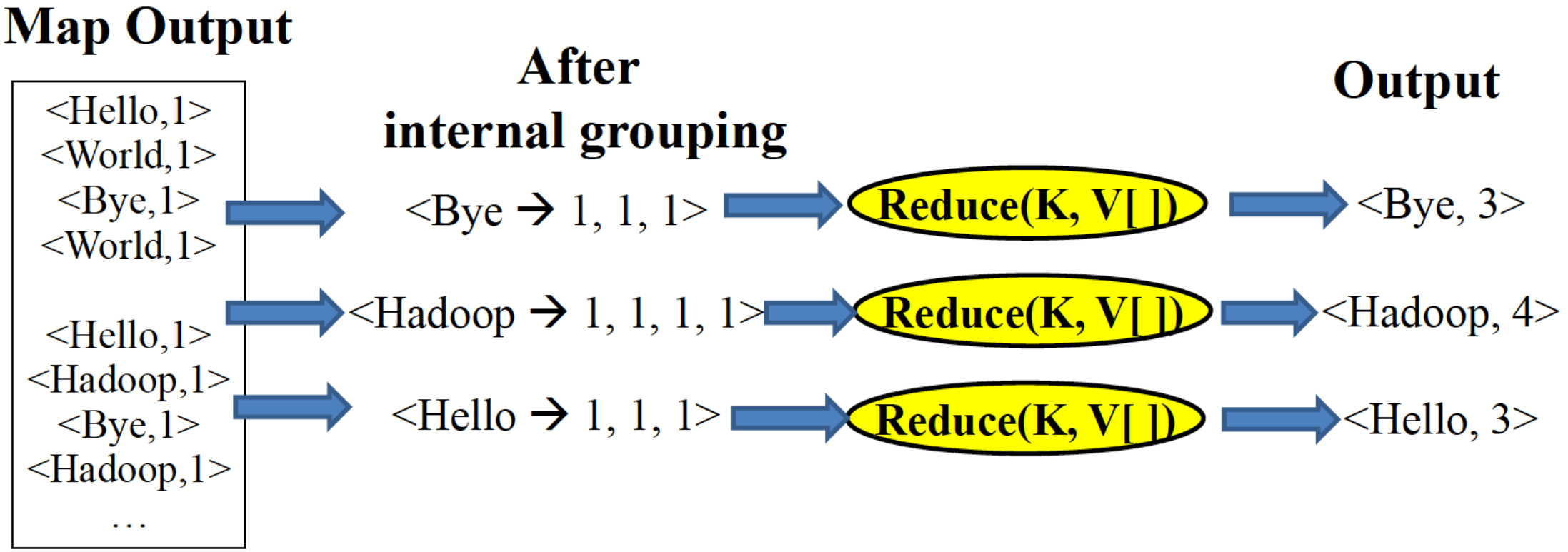
Word Counting: Map Function




```
Map(K, V)
{  For each word w in V,   Collect(w, 1);}
```

Slide adapted from Alexander Behm & Ajey Shah's presentation (http://www.slideshare.net/gothicane/behm-shah-pagerank)

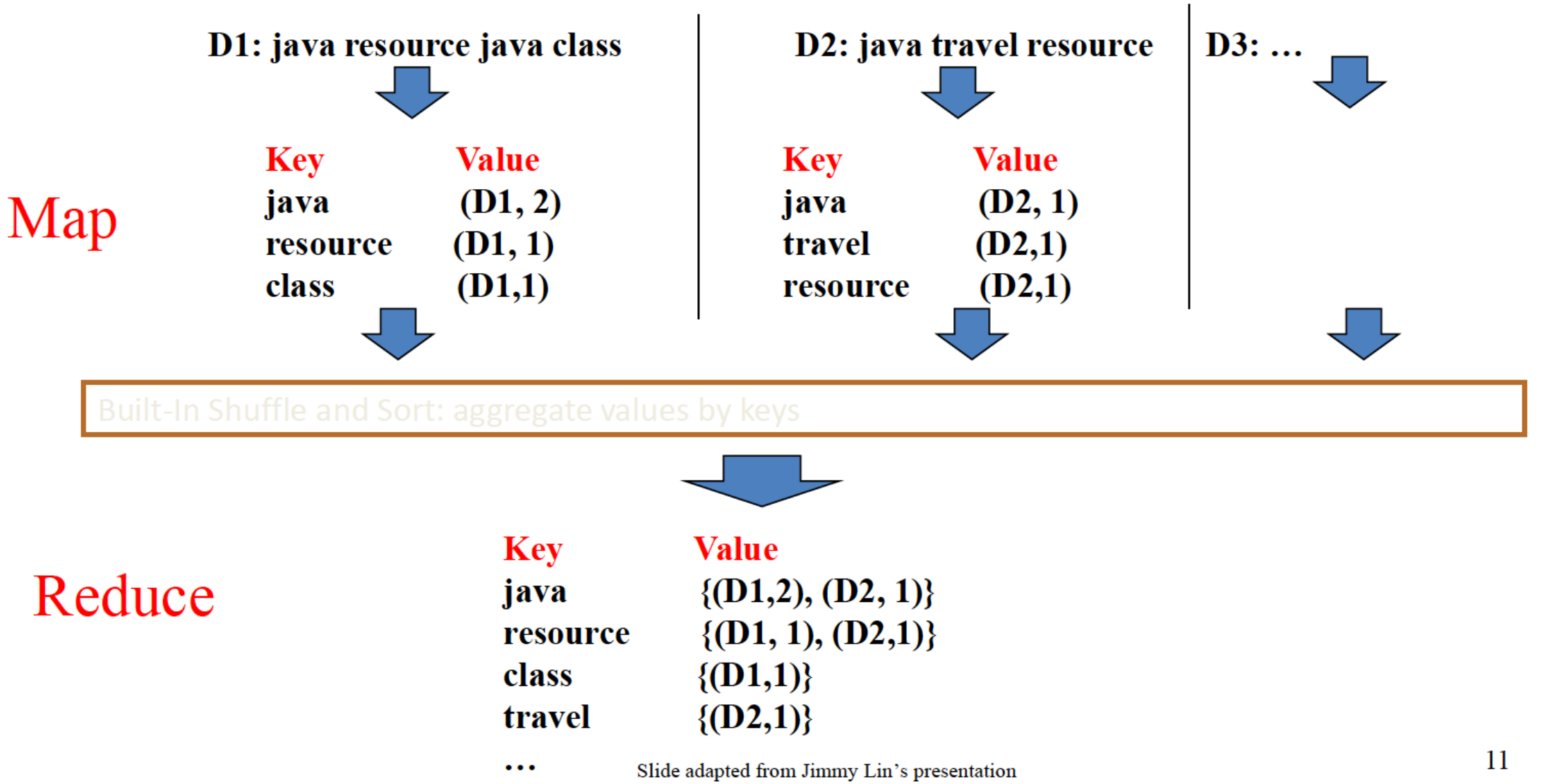
Word Counting: Reduce Function




```
Reduce(K, V[ ])
{ Int count = 0; For each v in V, count += v; Collect(K, count); }
```

Slide adapted from Alexander Behm & Ajey Shah's presentation (http://www.slideshare.net/gothicane/behm-shah-pagerank)

Inverted Indexing with MapReduce



Slide adapted from Jimmy Lin's presentation

Summary

- Web scale indexing requires
  - Storing the index on multiple machines (GFS)
  - Creating the index in parallel (MapReduce)
- Both GFS and MapReduce are general infrastructures