An Introduction

Kai Voigt, Cloudera Inc

Berlin Buzzwords, June 6th 2011

cloudera

Big Data

- Capture
- Storage
- Search
- Analytics

cloudera

hadoop.apache.org

Google	hadoop
Google Filesystem (GFS)	Hadoop Distributed Filesystem (HDFS)
MapReduce	MapReduce

cloudera

Hadoop Distributed Filesystem

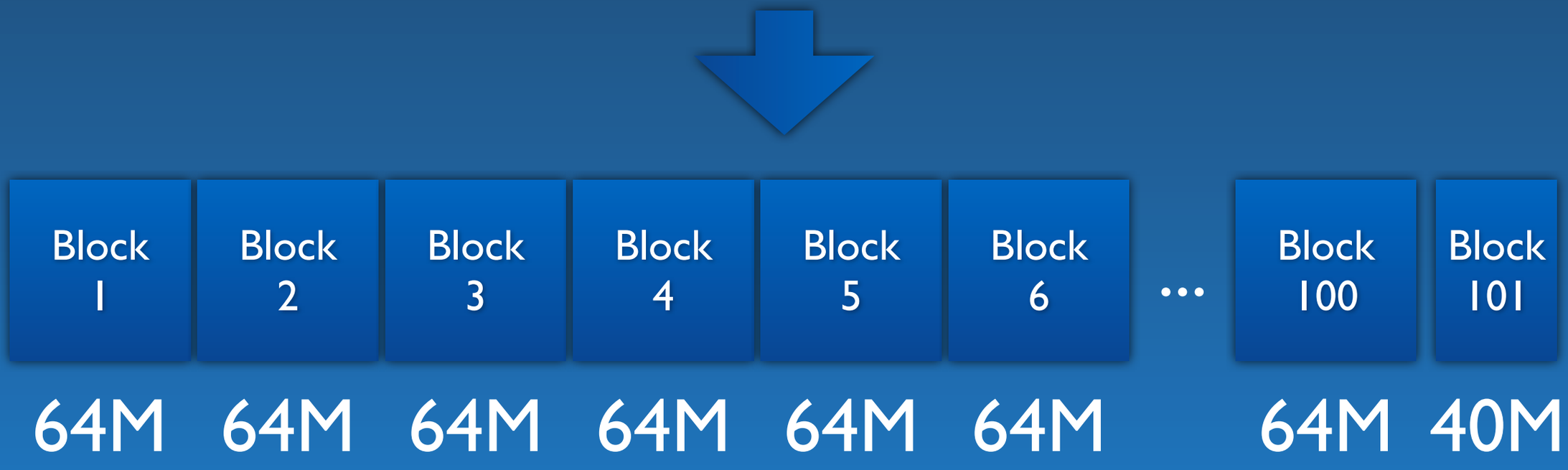
- Easy Access
- Distributed
- Redundant
- Scalable

cloudera

File Splits

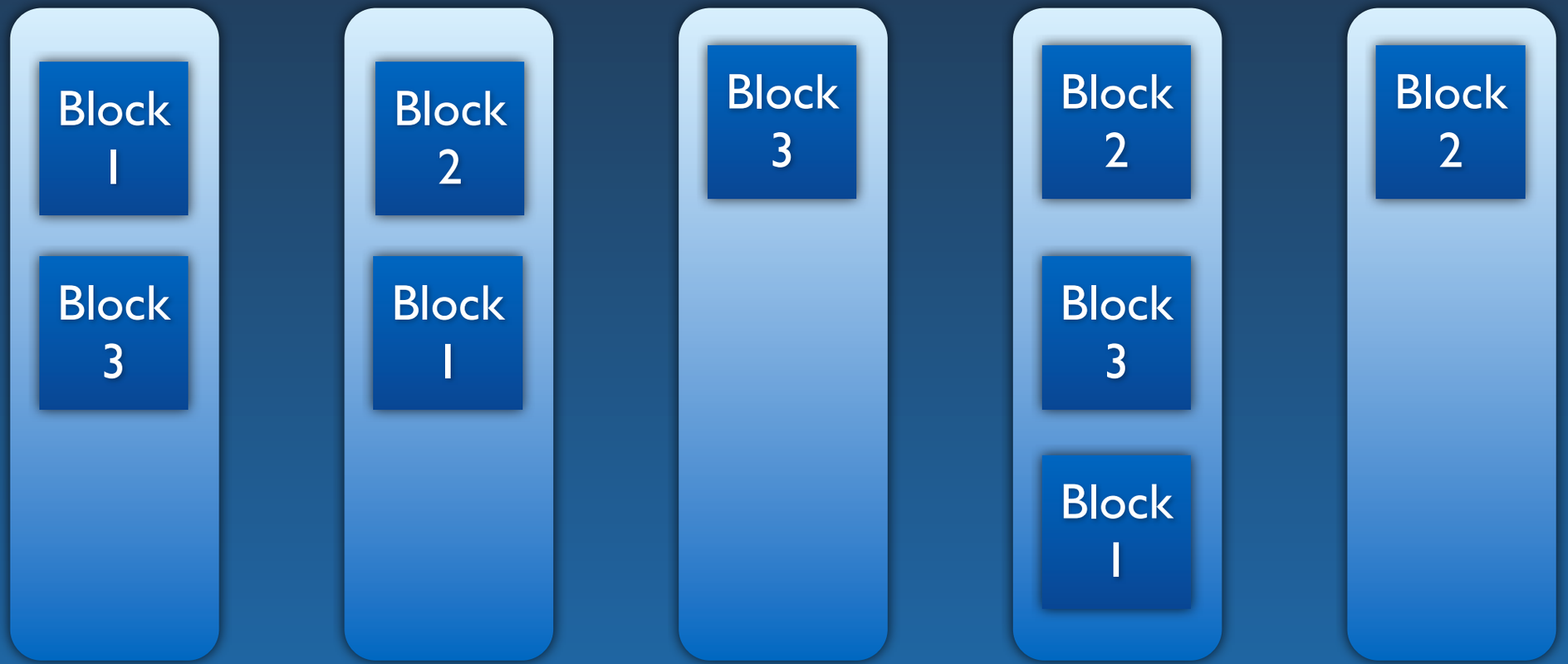
Large File

6440M



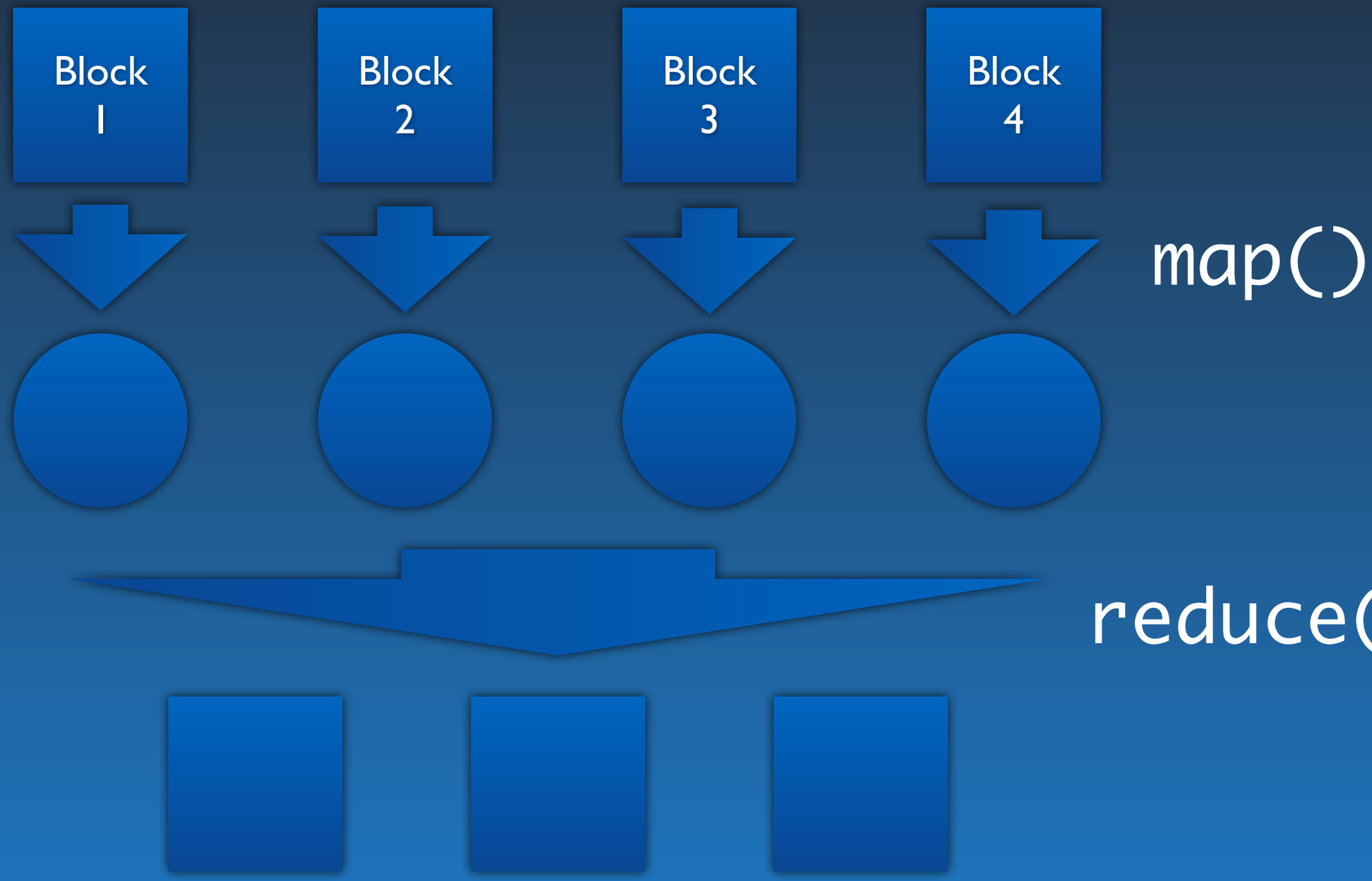
cloudera

Block Placement



cloudera

MapReduce



cloudera

WordCount Example

```
map (offset, line) {
   foreach word in line {
      emit (word, 1)
   }
}
```

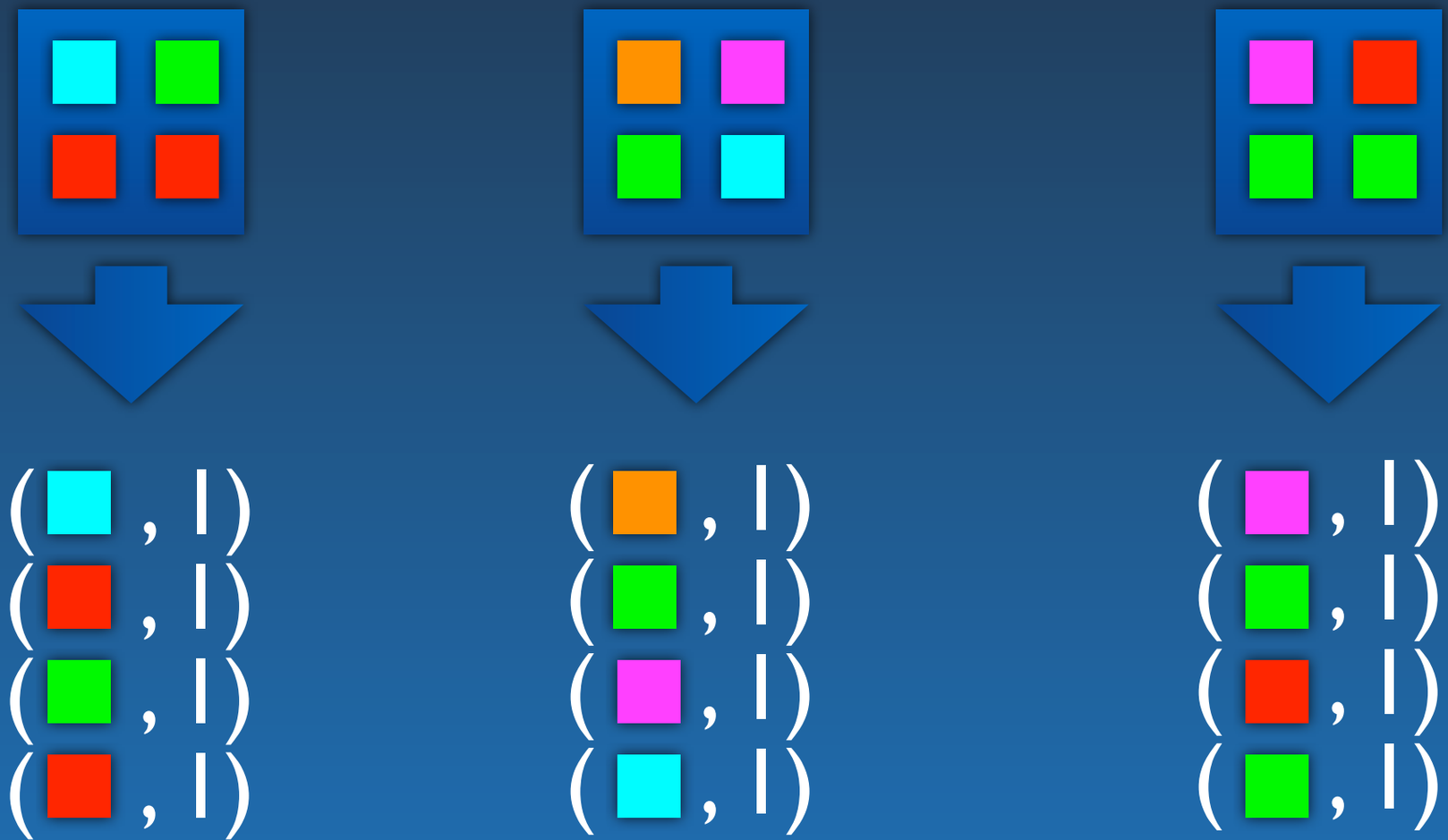
cloudera

WordCount Example

```
reduce (word, count[]) {
   total = 0;
   foreach number in count[] {
      total += number
   }
   emit (word, total)
}
```

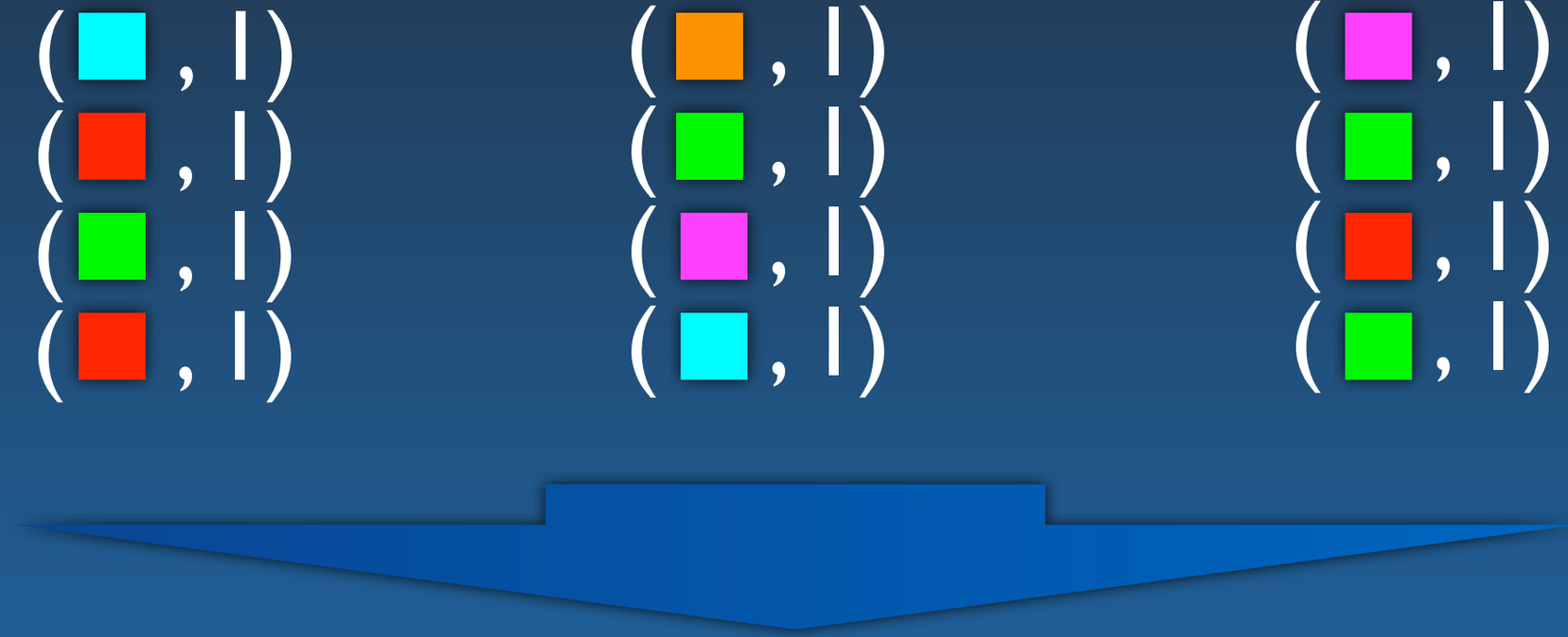
cloudera

map()



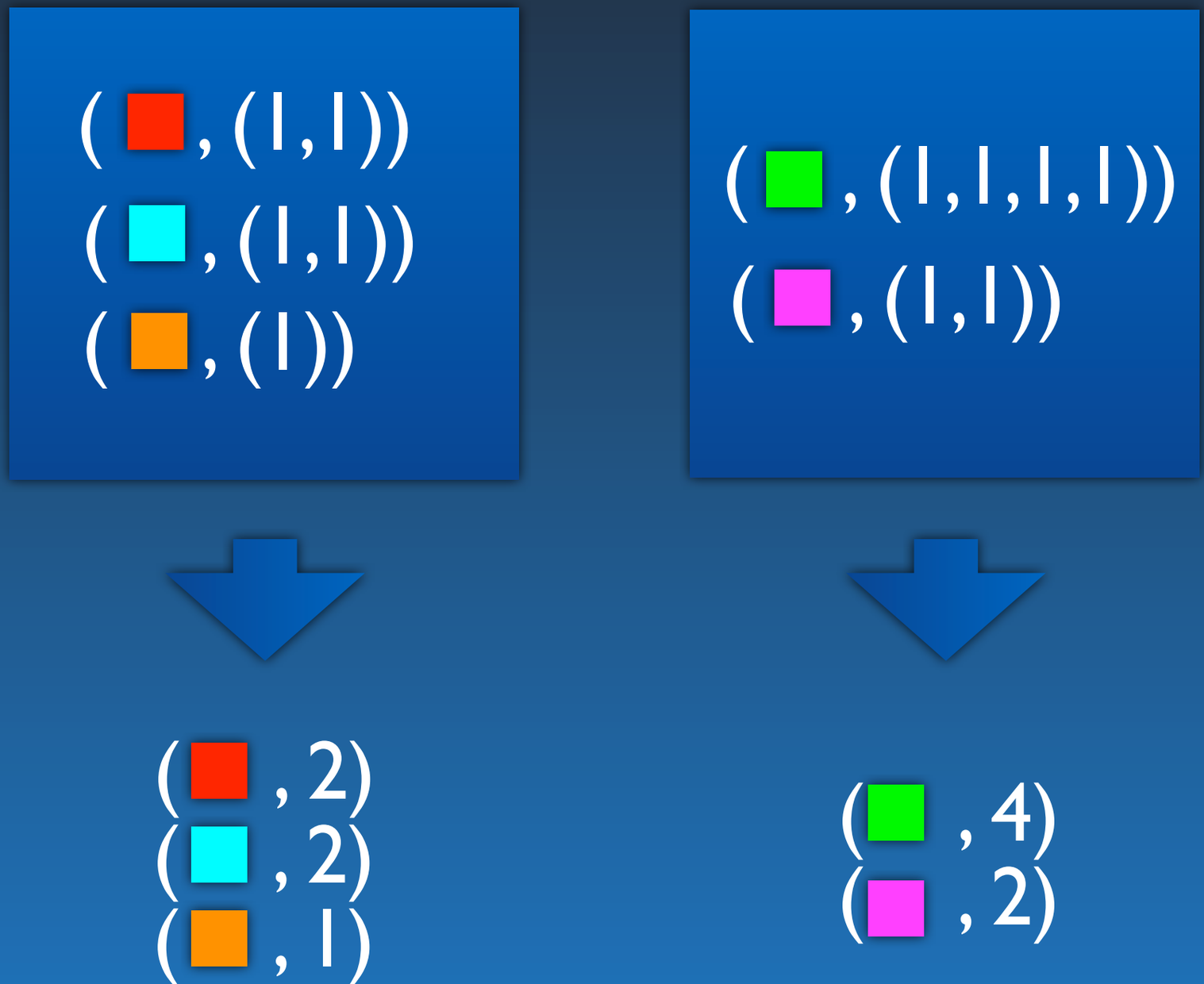
cloudera

Sort & Shuffle



cloudera

reduce()



cloudera

Use Case: Recommendations

- "People looking as this article also looked at these articles"
- "You might also know these people"
- iTunes Genius Playlist
- Banner Placement

cloudera

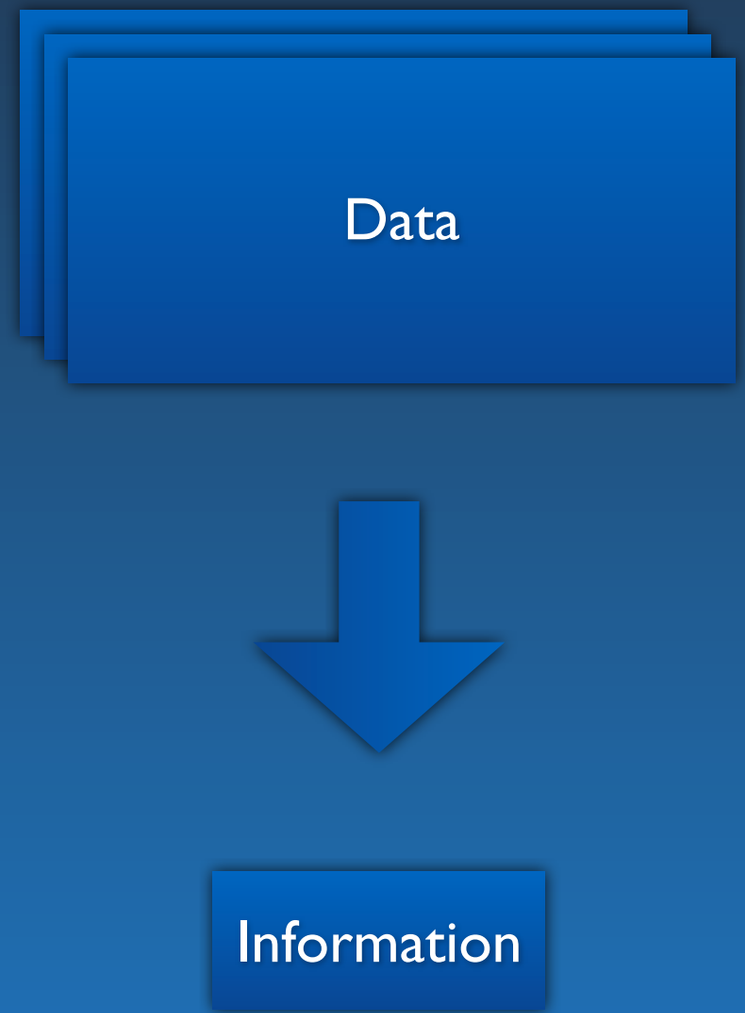
Use Case: Text Processing

- Document Indexing
- Semantic Analytics

cloudera

Use Case: Machine Learning

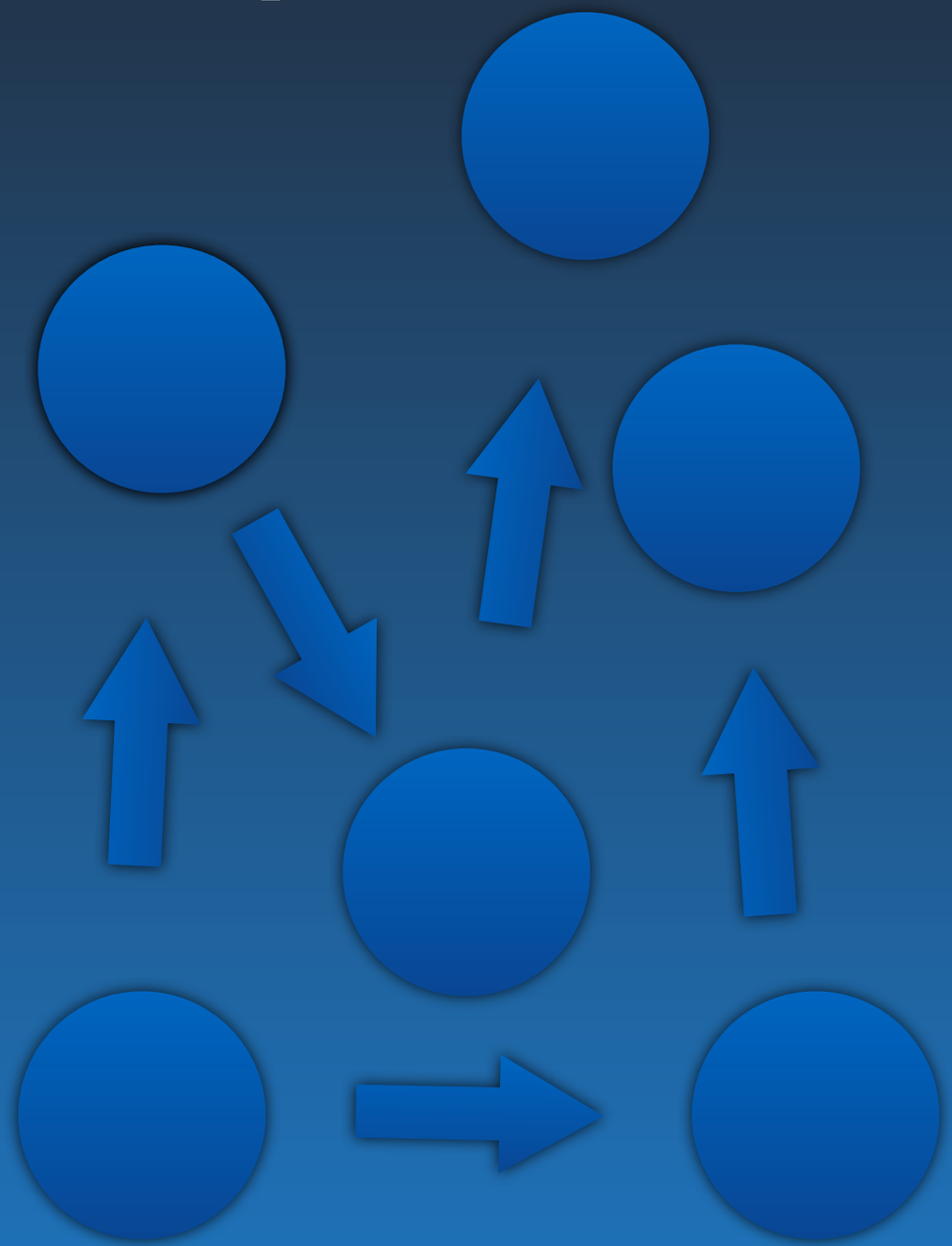
- Spam vs No Spam
- Credit Card Fraud
- "People of Interest"



cloudera

Use Case: Graphs

- Shortest Paths
- Bottleneck Nodes
- Flow Optimization
- Spanning Trees
- Spanning Routes



cloudera

Hadoop Ecosystem

Hive & Pig	High Level Language Access
HBase	Real Time Access
Sqoop	SQL to/from Hadoop
Flume	Distributed Data Collection
Oozie	Job Workflow
Mahout	Machine Learning Library

many more

cloudera

Homework

- Cloudera's Distribution including Hadoop (CDH3)
- Online Tutorials
- WordCount Example
- Conference Demo Cluster

cloudera

Quick Demo!

cloudera

Thank you!

- Kai Voigt
- kai@cloudera.com
- http://www.cloudera.com/
- http://apache.hadoop.org/

cloudera