

# Big Data Analytics

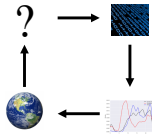
## Map-Reduce and Spark

### Review

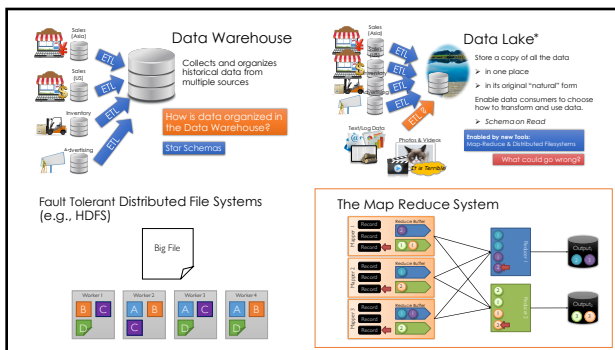
Slides by:

Joseph E. Gonzalez

[jgonzal@cs.berkeley.edu](mailto:jgonzal@cs.berkeley.edu)



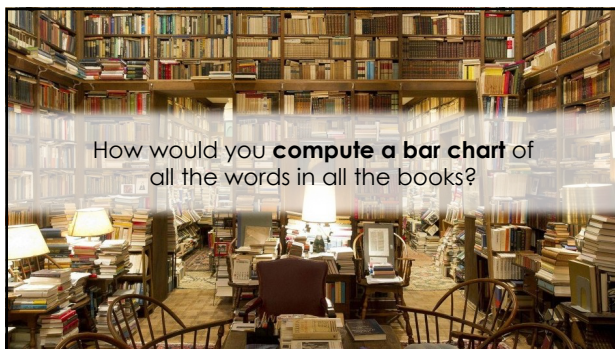
Previously ... previously ...



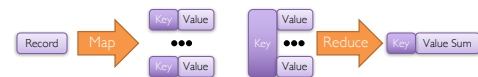
## Map-Reduce

### Distributed Aggregation

Computing are very large files



## The Map Reduce Abstraction



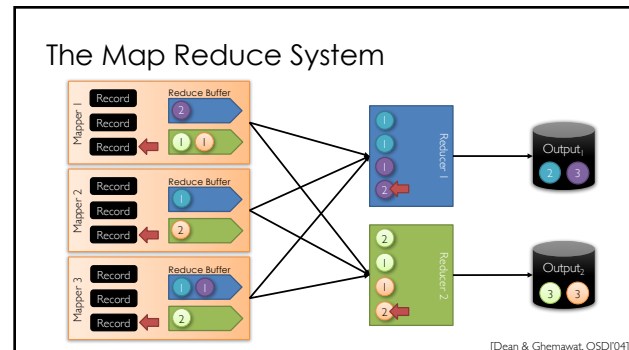
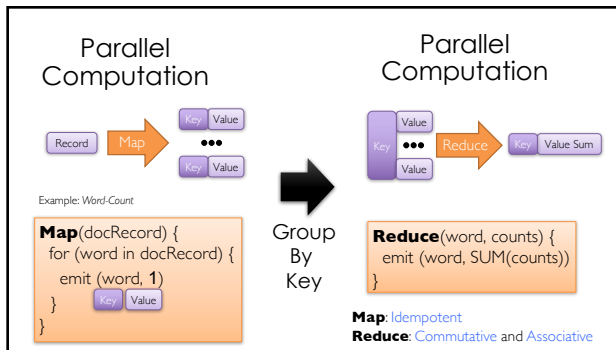
Example: Word-Count

```
Map(docRecord) {
  for (word in docRecord) {
    emit (word, 1)
  }
}
```

```
Reduce(word, counts) {
  emit (word, SUM(counts))
}
```

**Map:** Idempotent  
**Reduce:** Commutative and Associative

[Dean & Ghemawat, OSDI'04]



**Hadoop**

- First open-source map-reduce software system
  - Managed by Apache foundation
- Based on Google's
  - Map-Reduce
  - Google File System
- Several key technologies
  - **HDFS**: Hadoop File System
  - **Yarn**: Yet another resource negotiator
  - **MapReduce**: map-reduce compute framework
    - Difficult to use → increasingly less used

**Spark**

**In-Memory Dataflow System**

Developed at the UC Berkeley AMP Lab

M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: cluster computing with working sets. HotCloud 10  
 M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, I. Stoica. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. NSDI 2012

**Functional Programming API**

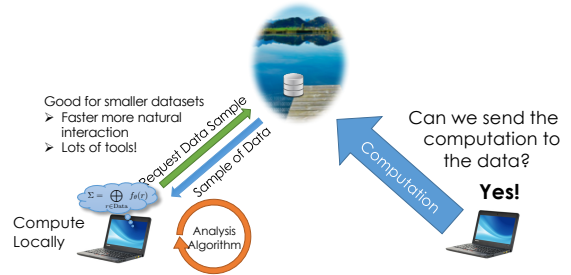
- Python, R, Java, and Scala APIs
- Data is represented as **distributed collections**:
  - **Resilient Distributed Dataset (RDD)**: A collection of **objects**
    - Objects could be text, images, python dictionaries ...
  - **DataFrame**: A collection of **rows** which all have the same schema (preferred way to use Spark today)
- Computation
  - **Transformations**: users apply functions to the distributed collections which return new distributed collections
    - E.g., map, filter, reduceByKey
  - **Actions**: users apply functions which return values:
    - E.g., count, sum, collect

**Multiple Demos!!**

# Interacting with Data @ Scale

Map-Reduce

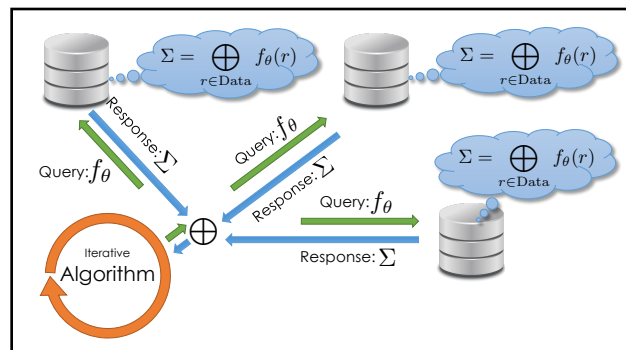
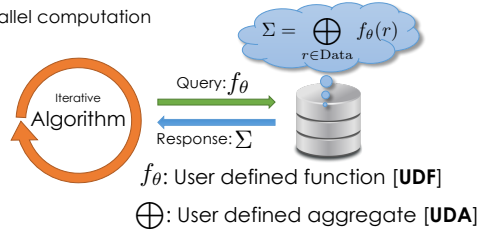
## Interacting With the Data



## Statistical Query Pattern

Common Machine Learning Pattern

- Computing aggregates of user defined functions
- Data-Parallel computation



## Interacting With the Data

