“Hadoop”: A Distributed Architecture, FileSystem, & MapReduce

Stony Brook University
CSE545 - Spring 2019
Big Data Analytics

- Workflow Systems
- Algorithms
- Statistical Methods
- Distributed Tools
Big Data Analytics

- Workflow Systems
- Algorithms
- Statistical Methods
- Distributed Tools
Classical Data Mining
Classical Data Mining
Classical Data Mining
Classical Data Mining
IO Bounded

Reading a word from disk versus main memory: $10^5$ slower!

Reading many contiguously stored words is faster per word, but fast modern disks still only reach 150MB/s for sequential reads.
IO Bounded

Reading a word from disk versus main memory: $10^5$ slower!

Reading many contiguously stored words is faster per word, but fast modern disks still only reach 150MB/s for sequential reads.

**IO Bound:** biggest performance bottleneck is reading / writing to disk.

---

starts around 100 GBs: ~10 minutes just to read

200 TBs: ~20,000 minutes = 13 days
Classical Big Data Analysis

Classical focus: efficient use of disk. e.g. Apache Lucene / Solr

Classical limitation: Still bounded when needing to process all of a large file.
IO Bound

How to solve?
Distributed Architecture (Cluster)
Distributed Architecture (Cluster)

In reality, modern setups often have multiple CPUs and disks per server, but we will model as if one machine per CPU-disk pair.
Distributed Architecture (Cluster)
Challenges for IO Cluster Computing

1. Nodes fail
   1 in 1000 nodes fail a day

2. Network is a bottleneck
   Typically 1-10 Gb/s throughput

3. Traditional distributed programming is often ad-hoc and complicated
Challenges for IO Cluster Computing

1. Nodes fail
   1 in 1000 nodes fail a day
   **Duplicate Data**

2. Network is a bottleneck
   Typically 1-10 Gb/s throughput
   **Bring computation to nodes, rather than data to nodes.**

3. Traditional distributed programming is often ad-hoc and complicated
   **Stipulate a programming system that can easily be distributed**
Challenges for IO Cluster Computing

1. Nodes fail
   1 in 1000 nodes fail a day
   **Duplicate Data**

2. Network is a bottleneck
   Typically 1-10 Gb/s throughput
   Bring computation to nodes, rather than data to nodes.

3. Traditional distributed programming is often ad-hoc and complicated
   Stipulate a programming system that can easily be distributed

**MapReduce Accomplishes**
Distributed File System

The effectiveness of MapReduce is in part simply due to use of a distributed filesystem!
Characteristics for Big Data Tasks

Large files (i.e. >100 GB to TBs)

Reads are most common

No need to update in place (append preferred)
Distributed File System

(e.g. Apache HadoopDFS, GoogleFS, EMRFS)

C, D: Two different files

(chunk server 1) (chunk server 2) (chunk server 3) ...

(chunk server n)

(Leskovec at al., 2014; http://www.mmds.org/)
Distributed File System

(e.g. Apache Hadoop DFS, GoogleFS, EM)

C, D: Two different files

“Hadoop” was named after a toy elephant belonging to Doug Cutting’s son. Cutting was one of Hadoop’s creators.

https://opensource.com/life/14/8/intro-apache-hadoop-big-data

(Leskovec at al., 2014; http://www.mmds.org/)
Distributed File System

(e.g. Apache Hadoop DFS, GoogleFS, EMRFS)

C, D: Two different files

(Leskovec at al., 2014; http://www.mmds.org/)
Distributed File System

(e.g. Apache HadoopDFS, GoogleFS, EMRFS)

C, D: Two different files

(chunk server 1) 

(chunk server 2) 

(chunk server 3) 

(chunk server n)

(Leskovec at al., 2014; http://www.mmds.org/)
Components of a Distributed File System

**Chunk servers (on Data Nodes)**

File is split into contiguous chunks
Typically each chunk is 16-64MB
Each chunk replicated (usually 2x or 3x)
Try to keep replicas in different racks

(Leskovec at al., 2014; http://www.mmds.org/)
Components of a Distributed File System

**Chunk servers (on Data Nodes)**
- File is split into contiguous chunks
- Typically each chunk is 16-64MB
- Each chunk replicated (usually 2x or 3x)
- Try to keep replicas in different racks

**Name node (aka master node)**
- Stores metadata about where files are stored
- Might be replicated or distributed across data nodes.

**Client library for file access**
- Talks to master to find chunk servers
- Connects directly to chunk servers to access data

(Leskovec at al., 2014; http://www.mmds.org/)
Challenges for IO Cluster Computing

1. Nodes fail
   1 in 1000 nodes fail a day
   *Duplicate Data (Distributed FS)*

2. Network is a bottleneck
   Typically 1-10 Gb/s throughput
   *Bring computation to nodes, rather than data to nodes.*

3. Traditional distributed programming is often ad-hoc and complicated
   *Stipulate a programming system that can easily be distributed*
What is MapReduce?

noun.1 - A style of programming

input chunks => map tasks  |  group_by keys  |  reduce tasks => output

"|" is the linux "pipe" symbol: passes stdout from first process to stdin of next.

E.g. counting words:

```
    tokenize(document)  |  sort  |  uniq -c
```
What is MapReduce?

*noun.1* - A *style of programming*

input chunks => **map tasks** | **group_by keys** | **reduce tasks** => output

“|” is the Linux “pipe” symbol: passes stdout from first process to stdin of next.

E.g. counting words:

```
 tokenize(document) | sort | uniq -c
```

*noun.2* - A *system* that distributes MapReduce style programs across a distributed file-system.

(e.g. Google’s internal “MapReduce” or apache.hadoop.mapreduce with hdfs)
What is MapReduce?
What is MapReduce?

input chunks

Key-value pairs (k, v)

Keys with all their values (k, [v, w, ...])

Group by keys

Reduce tasks

Combined output

Map

extract what you care about.

line => (k, v)
What is MapReduce?

Input chunks → Map → Sort and shuffle

Key-value pairs (k, v) => (k, [v, w, ...])

Keys with all their values

Map by keys → Reduce tasks

Combined output

extract what you care about.
What is MapReduce?

Input chunks → Map → sort and shuffle → Reduce → Combined output

- Extract what you care about.
- Aggregate, summarize.
What is MapReduce?

(Leskovec at al., 2014; http://www.mmds.org/)
The Map Step

Input key-value pairs

Intermediate key-value pairs

(Leskovec at al., 2014; http://www.mmds.org/)
The Sort / Group By Step

Intermediate key-value pairs

Key-value groups

Group by key

(Leskovec at al., 2014; http://www.mmds.org/)
The Reduce Step

(Leskovec at al., 2014; http://www.mmds.org/)
What is MapReduce?

(Leskovec at al., 2014; http://www.mmds.org/)
What is MapReduce?

Map: (k,v) -> (k’, v’)*
    (Written by programmer)

Group by key: (k_1’, v_1’), (k_2’, v_2’), ... -> (k_1’, (v_1’, v’, ...)),
    (system handles)               (k_2’, (v_1’, v’, ...), ...)

Reduce: (k’, (v_1’, v’, ...)) -> (k’, v’’)*
    (Written by programmer)
Example: Word Count

tokenize(document) | sort | uniq -C
Example: Word Count

tokenize(document) | sort | uniq -C

Map: extract what you care about.

sort and shuffle

Reduce: aggregate, summarize
Example: Word Count

The crew of the space shuttle Endeavor recently returned to Earth as ambassadors, harbingers of a new era of space exploration. Scientists at NASA are saying that the recent assembly of the Dextre bot is the first step in a long-term space-based man/machine partnership. "The work we're doing now -- the robotics we're doing -- is what we're going to need .................

Big document (Leskovec at al., 2014; http://www.mmds.org/)
The crew of the space shuttle Endeavor recently returned to Earth as ambassadors, harbingers of a new era of space exploration. Scientists at NASA are saying that the recent assembly of the Dextre bot is the first step in a long-term space-based man/machine partnership. "The work we're doing now -- the robotics we're doing -- is what we're going to need ......................

Big document (key, value)
The crew of the space shuttle Endeavor recently returned to Earth as ambassadors, harbingers of a new era of space exploration. Scientists at NASA are saying that the recent assembly of the Dextre bot is the first step in a long-term space-based man/machine partnership. "The work we're doing now -- the robotics we're doing -- is what we're going to need ...."
The crew of the space shuttle Endeavor recently returned to Earth as ambassadors, harbingers of a new era of space exploration. Scientists at NASA are saying that the recent assembly of the Dextre bot is the first step in a long-term space-based man/machine partnership.

"The work we're doing now -- the robotics we're doing -- is what we're going to need ....................

Big document

(key, value)

(key, value)

(key, value)
The crew of the space shuttle Endeavor recently returned to Earth as ambassadors, harbingers of a new era of space exploration. Scientists at NASA are saying that the recent assembly of the Dextre bot is the first step in a long-term space-based man/machine partnership. "The work we’re doing now -- the robotics we’re doing -- is what we’re going to need..."
Example: Word Count

```python
@abstractmethod
def map(k, v):
    pass

@abstractmethod
def reduce(k, vs):
    pass
```
Example: Word Count (version 1)

```python
def map(k, v):
    for w in tokenize(v):
        yield (w,1)

def reduce(k, vs):
    return len(vs)
```
Example: Word Count (version 1)

```python
def map(k, v):
    for w in tokenize(v):
        yield (w, 1)

def reduce(k, vs):
    return len(vs)
```

```python
def tokenize(s):
    # simple version
    return s.split(' ')
Example: Word Count (version 2)

```python
def map(k, v):
    counts = dict()
    for w in tokenize(v):
        try:
            counts[w] += 1
        except KeyError:
            counts[w] = 1
    for item in counts.items():
        yield item

def reduce(k, vs):
    return sum(vs)
```

- Counts each word within the chunk (try/except is faster than “if w in counts”)
- Sum of counts from different chunks
Challenges for IO Cluster Computing

1. Nodes fail
   1 in 1000 nodes fail a day
   Duplicate Data (Distributed FS)

2. Network is a bottleneck
   Typically 1-10 Gb/s throughput (Sort & Shuffle)
   Bring computation to nodes, rather than data to nodes.

3. Traditional distributed programming is often ad-hoc and complicated
   Stipulate a programming system that can easily be distributed
Challenges for IO Cluster Computing

1. Nodes fail
   1 in 1000 nodes fail a day
   **Duplicate Data** (Distributed FS)

2. Network is a bottleneck
   Typically 1-10 Gb/s throughput *(Sort & Shuffle)*
   Bring computation to nodes, rather than data to nodes.

3. Traditional distributed programming is often ad-hoc and complicated *(Simply requires Mapper and Reducer)*
   Stipulate a programming system that can easily be distributed
Example: Relational Algebra

Select
Project
Union, Intersection, Difference
Natural Join
Grouping
Example: Relational Algebra

Select

Project

Union, Intersection, Difference

Natural Join

Grouping
Example: Relational Algebra

**Select**

\[ R(A_1, A_2, A_3, \ldots) \], Relation \( R \), Attributes \( A_1 \),

return only those attribute tuples where condition \( C \) is true.
Example: Relational Algebra

**Select**

\[ R(A_1, A_2, A_3, \ldots), \text{ Relation } R, \text{ Attributes } A_* \]

return only those attribute tuples where condition \( C \) is true

```python
def map(k, v):
    # v is list of attribute tuples
    for t in v:
        if t satisfies C:
            yield (t, t)

def reduce(k, vs):
    For each v in vs:
        yield (k, v)
```
Example: Relational Algebra

**Natural Join**

Given $R_1$ and $R_2$ return $R_{\text{join}}$ -- union of all pairs of tuples that match given attributes.
Example: Relational Algebra

Natural Join

Given $R_1$ and $R_2$ return $R_{\text{join}}$ -- union of all pairs of tuples that match given attributes.

```python
def map(k, v):
    # k \in \{R1, R2\}, v is ($R_1$=(A, B), $R_2$=(B, C)); B are matched attributes
    if k=="R1":
        (a, b) = v
        yield (b,(R_1,a))
    if k=="R2":
        (b,c) = v
        yield (b,(R_2,c))
```
Example: Relational Algebra

Natural Join

Given $R_1$ and $R_2$ return $R_{\text{join}}$ -- union of all pairs of tuples that match given attributes.

```python
def map(k, v):
    # k \in \{R1, R2\}, v is (R_1=(A, B), R_2=(B, C)); B are matched attributes
    if k=="R1":
        (a, b) = v
        yield (b,(R_1,a))
    if k=="R2":
        (b,c) = v
        yield (b,(R_2,c))

def reduce(k, vs):
    r1, r2 = [], []
    for (S, x) in vs:
        # separate rs
        if S == r1: r1.append(x)
        else: r2.append(x)
    for a in r1:
        for each c in r2:
            yield (R_{\text{join}}, (a, k, c))
    #k is
```
Data Flow

MAP: Read input and produces a set of key-value pairs

Group by key: Collect all pairs with same key (Hash merge, Shuffle, Sort, Partition)

Reduce: Collect all values belonging to the key and output

Input: Big document

Intermediate:
- k1:v k1:v k2:v
- k1:v k3:v k4:v
- k4:v k5:v
- k4:v k1:v k3:v

Group by Key

Grouped:
- k1:v, v, v, v
- k2:v
- k3:v, v
- k4:v, v, v, v
- k5:v

Output

Data Flow: In Parallel

(Map Task 1)

Map Task 1

k1:v k1:v k2:v k1:v
Partitioning Function

Map Task 2

k3:v k4:v k4:v k5:v
Partitioning Function

Map Task 3

k4:v k1:v k3:v
Partitioning Function

Sort and Group

k2:v k4:v k5:v

Reduce Task 1

(R R R)

Sort and Group

k1:v k3:v

Reduce Task 2

(R R)

(hash)

(Leskovec at al., 2014; http://www.mmds.org/)
Data Flow: In Parallel

(Leskovec at al., 2014; http://www.mmds.org/)
Data Flow

DFS ➔ Map ➔ Map’s Local FS ➔ Reduce ➔ DFS
Data Flow

MapReduce system handles:

- Partitioning
- Scheduling map / reducer execution
- Group by key
- Restarts from node failures
- Inter-machine communication
Data Flow

- Schedule map tasks near physical storage of chunk
- Intermediate results stored locally
- Master / Name Node coordinates
Data Flow

- Schedule map tasks near physical storage of chunk
- Intermediate results stored locally
- Master / Name Node coordinates
  - Task status: idle, in-progress, complete
  - Receives location of intermediate results and schedules with reducer
  - Checks nodes for failures and restarts when necessary
    - All map tasks on nodes must be completely restarted
    - Reduce tasks can pickup with reduce task failed
Data Flow

- Schedule map tasks near physical storage of chunk
- Intermediate results stored locally
- Master / Name Node coordinates
  - Task status: idle, in-progress, complete
  - Receives location of intermediate results and schedules with reducer
  - Checks nodes for failures and restarts when necessary
    - All map tasks on nodes must be completely restarted
    - Reduce tasks can pickup with reduce task failed

DFS → MapReduce → DFS → MapReduce → DFS
Data Flow

Skew: The degree to which certain tasks end up taking much longer than others.

Handled with:

- More reducers than reduce tasks
- More reduce tasks than nodes
Data Flow

**Key Question:** How many Map and Reduce jobs?
Data Flow

Key Question: How many Map and Reduce jobs?

$M$: map tasks, $R$: reducer tasks

$A$: If possible, one chunk per map task

and $M \gg |\text{nodes}| \approx |\text{cores}|$

(better handling of node failures, better load balancing)

$R < M$

(reduces number of parts stored in DFS)
Data Flow

version 1: few reduce tasks
(same number of reduce tasks as nodes)

Reduce tasks represented by
time to complete task
(some tasks take much longer)
Data Flow

version 1: few reduce tasks
(same number of reduce tasks as nodes)

Reduce tasks represented by
**time to complete task**
(some tasks take much longer)

version 2: more reduce tasks
(more reduce tasks than nodes)

Reduce tasks represented by
**time to complete task**
(some tasks take much longer)
Data Flow

version 1: few reduce tasks
(same number of reduce tasks as nodes)

Reduce tasks represented by
\textbf{time to complete task}
(some tasks take much longer)

Reduce tasks represented by
\textbf{time to complete task}
(some tasks take much longer)

version 2: more reduce tasks
(more reduce tasks than nodes)

Reduce tasks represented by
\textbf{time to complete task}
(some tasks take much longer)

\textbf{Last task completed}

Can redistribute these tasks to other nodes

\textbf{(the last task now completes much earlier)}
Communication Cost Model

How to assess performance?

(1) Computation: Map + Reduce + System Tasks

(2) Communication: Moving (key, value) pairs
Communication Cost Model

How to assess performance?

1. Computation: Map + Reduce + System Tasks

2. Communication: Moving (key, value) pairs

Ultimate Goal: wall-clock Time.
Communication Cost Model

How to assess performance?

(1) Computation: Map + Reduce + System Tasks

- Mappers and reducers often single pass $O(n)$ within node
- System: sort the keys is usually most expensive
- Even if map executes on same node, disk read usually dominates
- In any case, can add more nodes

(2) Communication: Moving key-value pairs

Ultimate Goal: wall-clock time.
Communication Cost Model

**How to assess performance?**

1. Computation: Map + Reduce + System Tasks

2. Communication: Moving key, value pairs

   Often dominates computation.
   - Connection speeds: 1-10 gigabits per sec;
   - HD read: 50-150 gigabytes per sec
   - Even reading from disk to memory typically takes longer than operating on the data.
Communication Cost Model

How to assess performance?

Communication Cost = input size + (sum of size of all map-to-reducer files)

(2) Communication: Moving key, value pairs

Often dominates computation.

- Connection speeds: 1-10 gigabits per sec;
  HD read: 50-150 gigabytes per sec
- Even reading from disk to memory typically takes longer than operating on the data.
Communication Cost Model

How to assess performance?

\[
\text{Communication Cost} = \text{input size} + \text{(sum of size of all map-to-reducer files)}
\]

(2) Communication: Moving key, value pairs

Often dominates computation.

- Connection speeds: 1-10 gigabits per sec;
  HD read: 50-150 gigabytes per sec
- Even reading from disk to memory typically takes longer than operating on the data.
- Output from reducer ignored because it’s either small (finished summarizing data) or being passed to another mapreduce job.
Example: Natural Join

$R, S$: Relations (Tables) \quad R(A, B) \bowtie S(B, C)$

**Communication Cost** = input size + (sum of size of all map-to-reducer files)

DFS $\rightarrow$ Map $\rightarrow$ LocalFS $\rightarrow$ Network $\rightarrow$ Reduce $\rightarrow$ DFS $\rightarrow$ ?
Example: Natural Join

R, S: Relations (Tables) \[ R(A, B) \bowtie S(B, C) \]

Communication Cost = input size + (sum of size of all map-to-reducer files)

def map(k, v):
    if k==”R1”:
        (a, b) = v
        yield (b,(R_1,a))
    if k==”R2”:
        (b,c) = v
        yield (b,(R_2,c))

def reduce(k, vs):
    r1, r2 = [], []
    for (rel, x) in vs: #separate rs
        if rel == ‘R’: r1.append(x)
        else: r2.append(x)
    for a in r1: #join as tuple
        for each c in r2:
            yield (R_{join}, (a, k, c)) #k is
Example: Natural Join

R, S: Relations (Tables) \[ R(A, B) \bowtie S(B, C) \]

**Communication Cost =** input size + (sum of size of all map-to-reducer files)

= \(|R_1| + |R_2| + (|R_1| + |R_2|)\)

= \(O(|R_1| + |R_2|)\)

```python
def map(k, v):
    if k=="R1":
        (a, b) = v
        yield (b,(R_1,a))
    if k=="R2":
        (b,c) = v
        yield (b,(R_2,c))

def reduce(k, vs):
    r1, r2 = [], []
    for (rel, x) in vs: #separate rs
        if rel == ‘R’: r1.append(x)
        else: r2.append(x)
    for a in r1: #join as tuple
        for each c in r2:
            yield (R_join,, (a, k, c)) #k is
Exercise:

Calculate Communication Cost for
“Matrix Multiplication with One MapReduce Step”
(see MMDS section 2.3.10)
Last Notes: Further Considerations for MapReduce

- **Performance Refinements:**
  - **Backup tasks (aka speculative tasks)**
    - Schedule multiple copies of tasks when close to the end to mitigate certain nodes running slow.
  - **Combiners (like word count version 2 but done via reduce)**
    - Run reduce right after map from same node before passing to reduce
    - Reduces communication cost
  - **Override partition hash function**
    - E.g. instead of hash(url) use hash(hostname(url))