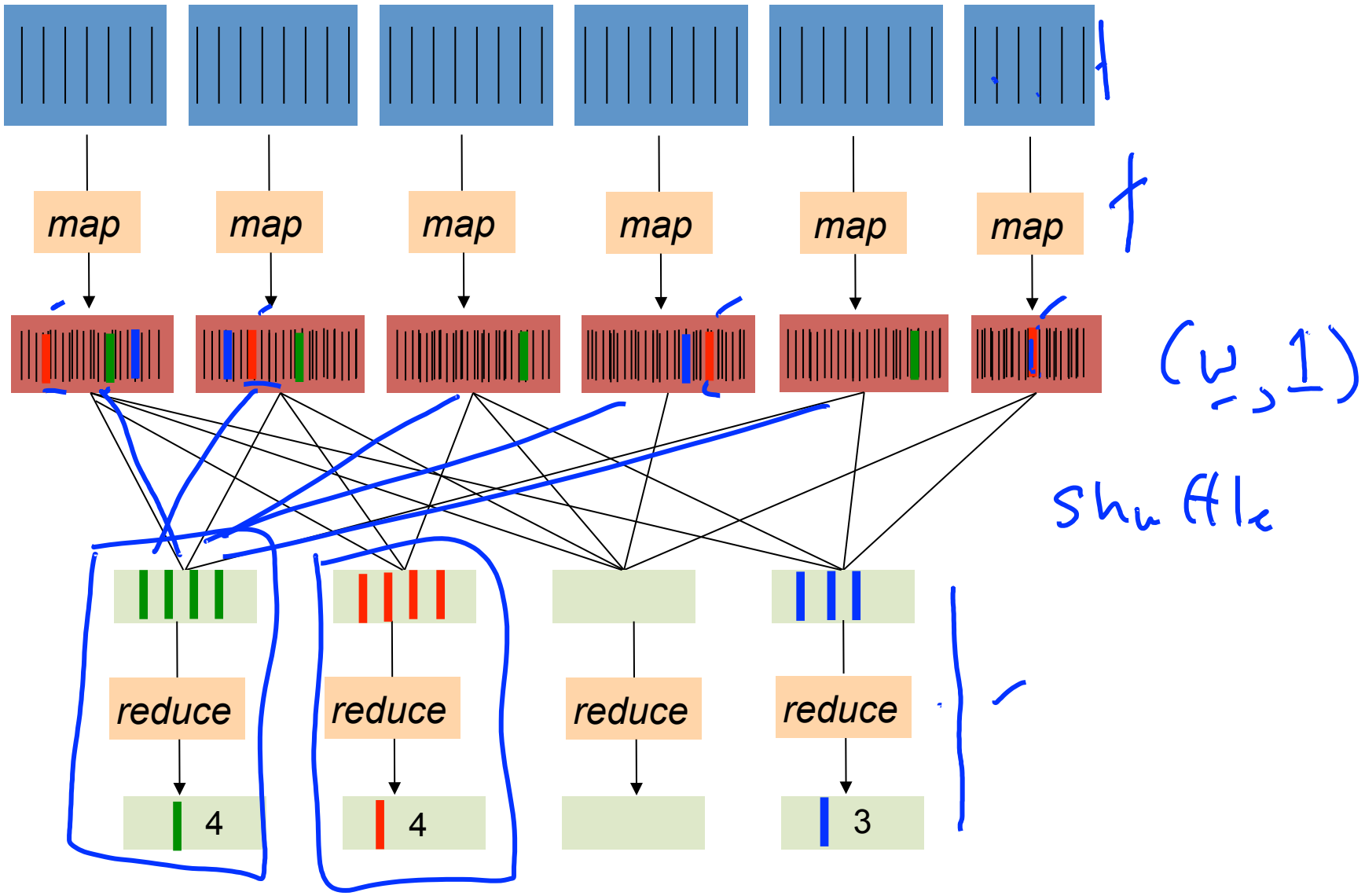
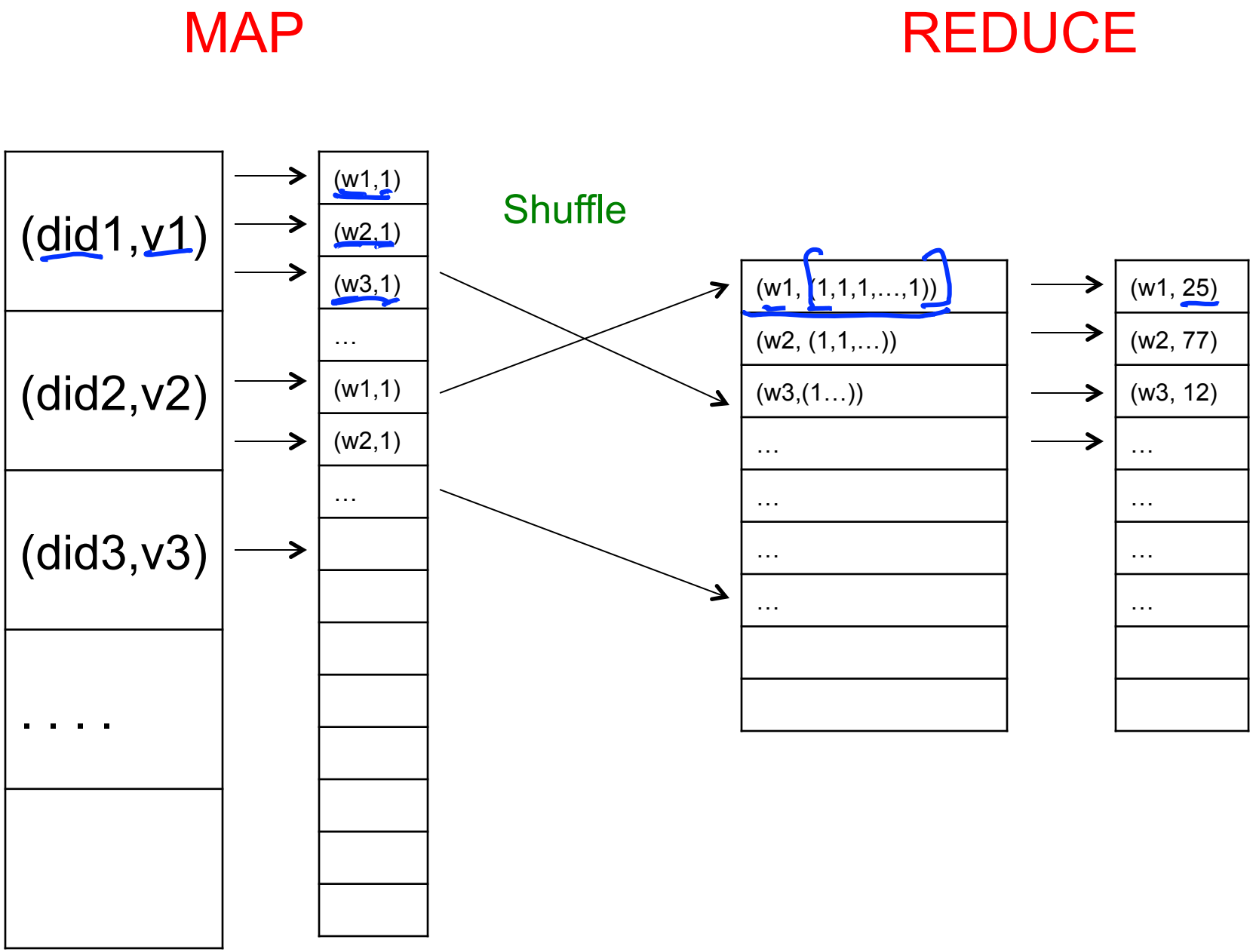
Count word occurrences across all documents




MAP
REDUCE
(did1,v1)
(did2,v2)
(did3,v3)
. . . .
(w1,1)
(w2,1)
(w3,1)
…
(w1,1)
(w2,1)
…
Shuffle
(w1, (1,1,1,…,1))
(w2, (1,1,…))
(w3,(1…))
…
…
…
…
(w1, 25)
(w2, 77)
(w3, 12)
…
…
…
…

Map Reduce

- Google: paper published 2004
- Free variant: Hadoop

- Map-reduce = high-level programming model and implementation for large-scale parallel data processing

Data Model

Files !

A file = a bag of **(key, value)** pairs

A map-reduce program:

- Input: a bag of **(inputkey, value)** pairs
- Output: a bag of **(outputkey, value)** pairs

Step 1: the MAP Phase

User provides the MAP-function:

- Input: **`(input key, value)`**
- Ouput:
  bag of **`(intermediate key, value)`**

System applies the map function in parallel to all **`(input key, value)`** pairs in the input file

Step 2: the REDUCE Phase

User provides the REDUCE function:

- Input:
  `(intermediate key, bag of values)`
- Output: bag of output **`(values)`**

The system will group all pairs with the same intermediate key, and passes the bag of values to the REDUCE function

MapReduce Programming Model

- Input & Output: each a set of key/value pairs
- Programmer specifies two functions:

map (in_key, in_value) -> list(out_key, intermediate_value)

- Processes input key/value pair
- Produces set of intermediate pairs

reduce (out_key, list(intermediate_value)) -> list(out_value)

- Combines all intermediate values for a particular key
- Produces a set of merged output values (usually just one)

Inspired by primitives from functional programming languages such as Lisp, Scheme, and Haskell

slide source: Google, Inc.

Example: What does this do?

```
map(String input_key, String input_value):
  // input_key: document name
  // input_value: document contents
  for each word w in input_value:
      EmitIntermediate(w, 1);


reduce(String output_key, Iterator intermediate_values):
  // output_key: word
  // output_values: ????
  int result = 0;
  for each v in intermediate_values:
      result += v;
  Emit(result);
```

slide source: Google, Inc.

Example: Word length histogram

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

How many “big”, “medium”, and “small” words are used?

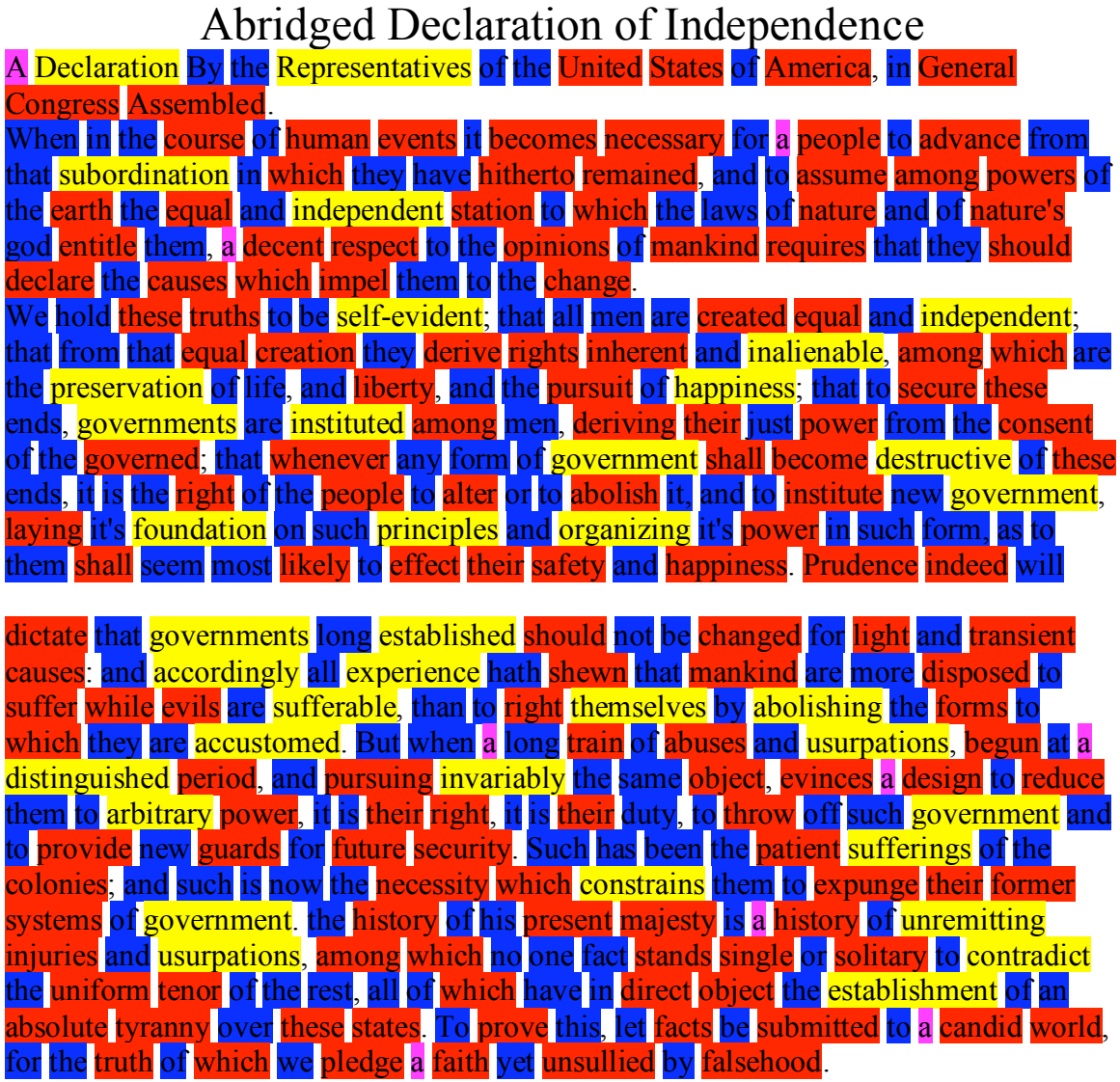
Example: Word length histogram

Big = Yellow = 10+ letters

Medium = Red = 5..9 letters

Small = Blue = 2..4 letters

Tiny = Pink = 1 letter



Example: Word length histogram

Split the document into chunks and process each chunk on a different computer

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

Chunk 1

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

Chunk 2

Example: Word length histogram

Abridged Declaration of Independence

Map Task 1 (204 words)

A Declaration By the Representatives of the United States of America, in General Congress Assembled.
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

(key, value)

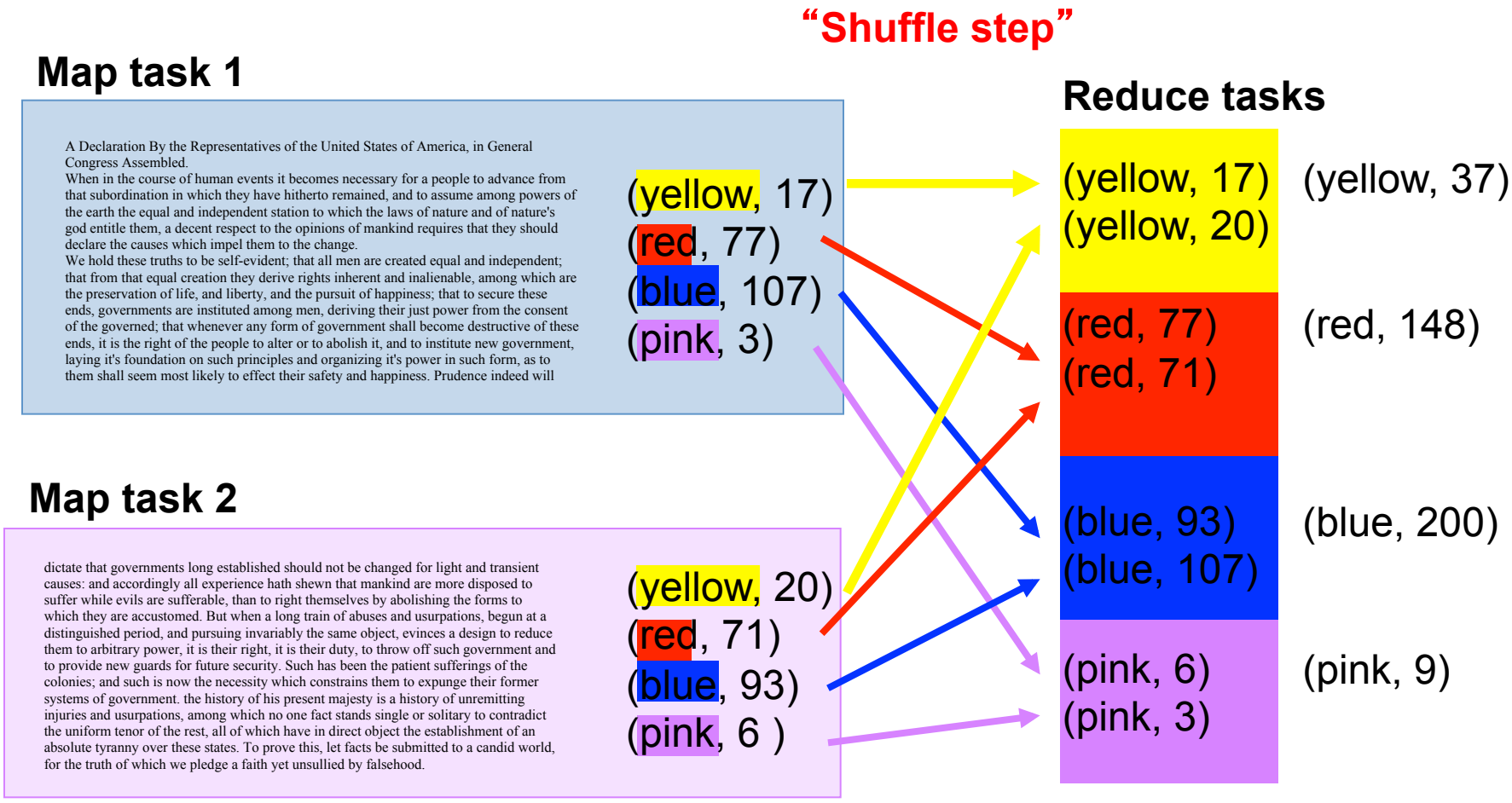
(yellow, 17)
(red, 77)
(blue, 107)
(pink, 3)

Map Task 2 (190 words)

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

(yellow, 20)
(red, 71)
(blue, 93)
(pink, 6)

Example: Word length histogram



More Examples: Build an Inverted Index

Input:

tweet1, (“I love pancakes for breakfast”)
tweet2, (“I dislike pancakes”)
tweet3, (“What should I eat for breakfast?”)
tweet4, (“I love to eat”)

Desired output:

“pancakes”, (tweet1, tweet2)
“breakfast”, (tweet1, tweet3)
“eat”, (tweet3, tweet4)
“love”, (tweet1, tweet4)
…

More Examples: Relational Join

Employee

Name	SSN
Sue	999999999
Tony	777777777

Assigned Departments

EmpSSN	DepName
999999999	Accounts
777777777	Sales
777777777	Marketing

Emplyee ⋈ Assigned Departments

Name	SSN	EmpSSN	DepName
Sue	999999999	999999999	Accounts
Tony	777777777	777777777	Sales
Tony	777777777	777777777	Marketing

Relational Join in MapReduce: Before Map Phase

Key idea: Lump all the tuples together into one dataset

Employee

Name	SSN
Sue	999999999
Tony	777777777

Assigned Departments

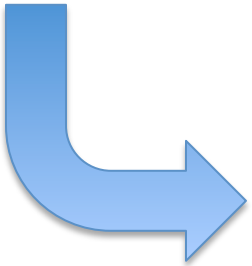
EmpSSN	DepName
999999999	Accounts
777777777	Sales
777777777	Marketing

Employee, Sue, 999999999
Employee, Tony, 777777777
Department, 999999999, Accounts
Department, 777777777, Sales
Department, 777777777, Marketing

What is this for?

Relational Join in MapReduce: Map Phase

Employee, Sue, 999999999
Employee, Tony, 777777777
Department, 999999999, Accounts
Department, 777777777, Sales
Department, 777777777, Marketing



key=999999999, value=(Employee, Sue, 999999999)
key=777777777, value=(Employee, Tony, 777777777)
key=999999999, value=(Department, 999999999, Accounts)
key=777777777, value=(Department, 777777777, Sales)
key=777777777, value=(Department, 777777777, Marketing)

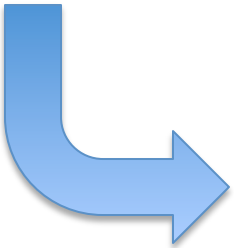
why do we use this as the key?

Relational Join in MapReduce: Reduce Phase

key=999999999, values=[(Employee, Sue, 999999999),
(Department, 999999999, Accounts)]

Sue, 999999999, 999999999, Accounts

key=777777777, values=[(Employee, Tony, 777777777),
(Department, 777777777, Sales),
(Department, 777777777, Marketing)]



Tony, 777777777, 777777777, Sales
Tony, 777777777, 777777777, Marketing

Relational Join in MapReduce, again

Order(orderid, account, date)

1, aaa, d1
2, aaa, d2
3, bbb, d3

LineItem(orderid, itemid, qty)

1, 10, 1
1, 20, 3
2, 10, 5
2, 50, 100
3, 20, 1

Map

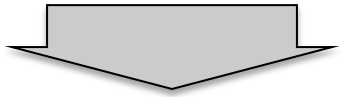
tagged with relation name

Order
- 1, aaa, d1 → 1 : "Order", (1,aaa,d1)
- 2, aaa, d2 → 2 : "Order", (2,aaa,d2)
- 3, bbb, d3 → 3 : "Order", (3,bbb,d3)

Line
- 1, 10, 1 → 1 : "Line", (1, 10, 1)
- 1, 20, 3 → 1 : "Line", (1, 20, 3)
- 2, 10, 5 → 2 : "Line", (2, 10, 5)
- 2, 50, 100 → 2 : "Line", (2, 50, 100)
- 3, 20, 1 → 3 : "Line", (3, 20, 1)

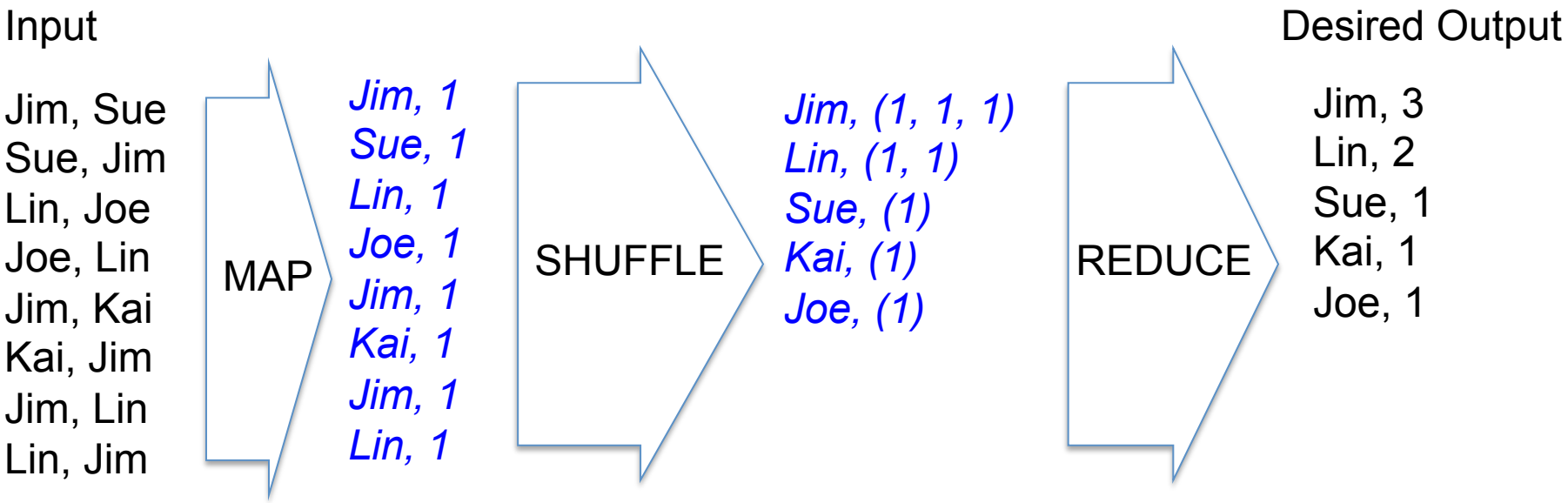
Reducer for key 1

"Order", (1,aaa,d1)
"Line", (1, 10, 1)
"Line", (1, 20, 3)



(1, aaa, d1, 1, 10, 1)
(1, aaa, d1, 1, 20, 3)

Simple Social Network Analysis: Count Friends



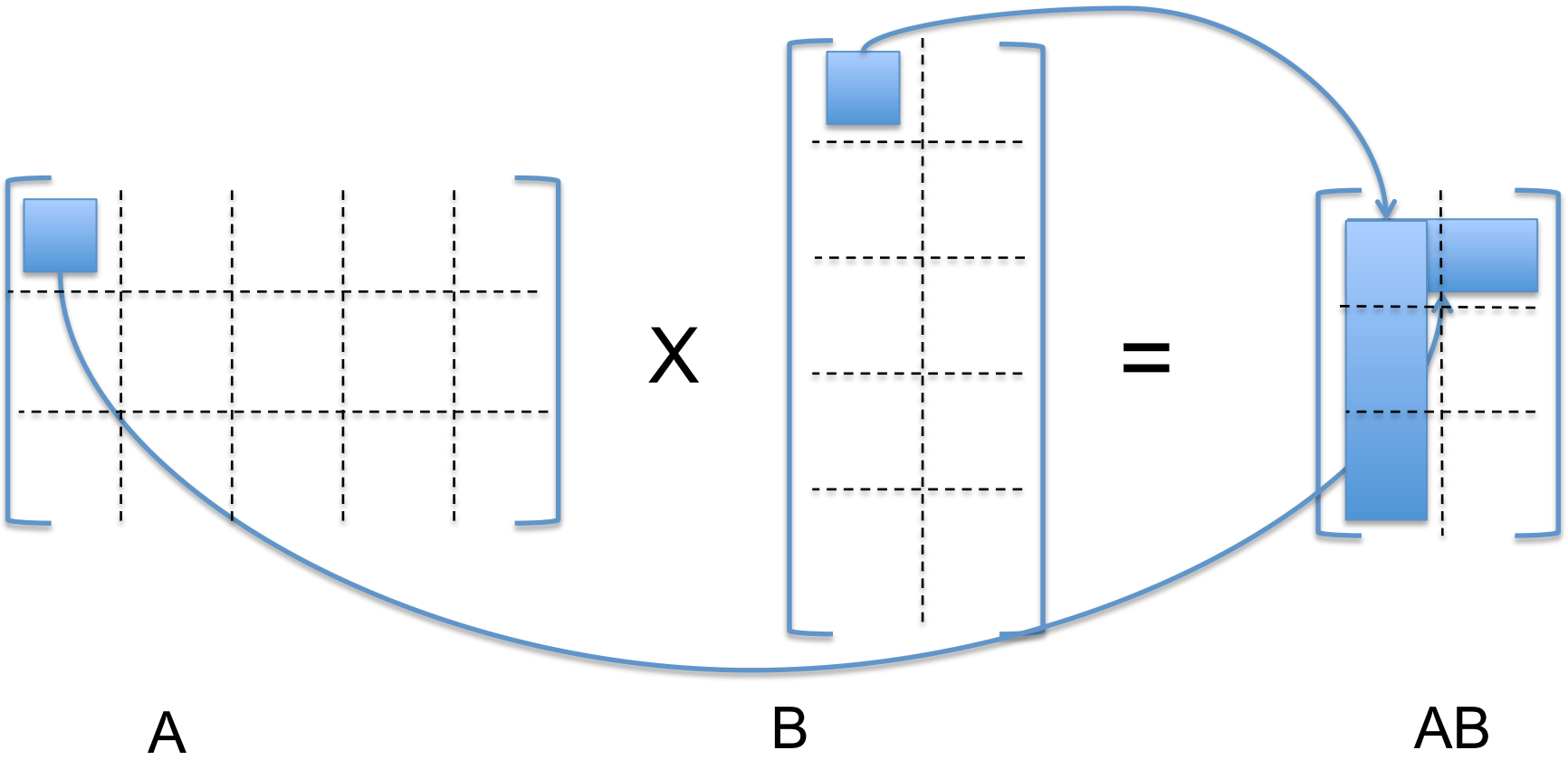
Matrix Multiply in MapReduce

C = A X B

A has dimensions L,M

B has dimensions M,N

- In the map phase:
  - for each element (i,j) of A, emit ((i,k), A[i,j]) for k in 1..N
  - for each element (j,k) of B, emit ((i,k), B[j,k]) for i in 1..L
- In the reduce phase, emit
  - key = (i,k)
  - value = Sum_j (A[i,j] * B[j,k])



- One reducer per output cell
- Each reducer computes Sum_j (A[i,j] * B[j,k])

Cluster Computing

- Large number of commodity servers, connected by high speed, commodity network
- Rack: holds a small number of servers
- Data center: holds many racks

READING ASSIGNMENT:
Map-reduce (Section 20.2); Chapter 2 (Sections 1,2,3 only)
of Mining of Massive Datasets, by Rajaraman and Ullman
See the course Calendar Website

Cluster Computing

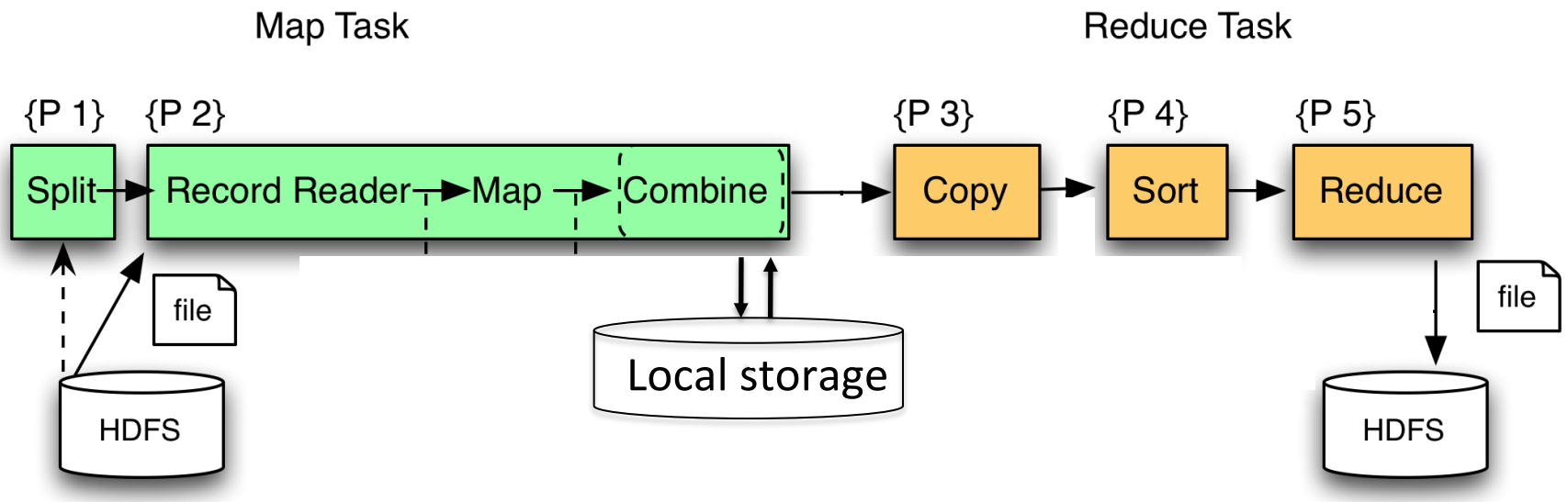
- Massive parallelism:
  - 100s, or 1000s, or 10000s servers
  - Many hours
- Failure:
  - If medium-time-between-failure is 1 year
  - Then 10000 servers have one failure / hour

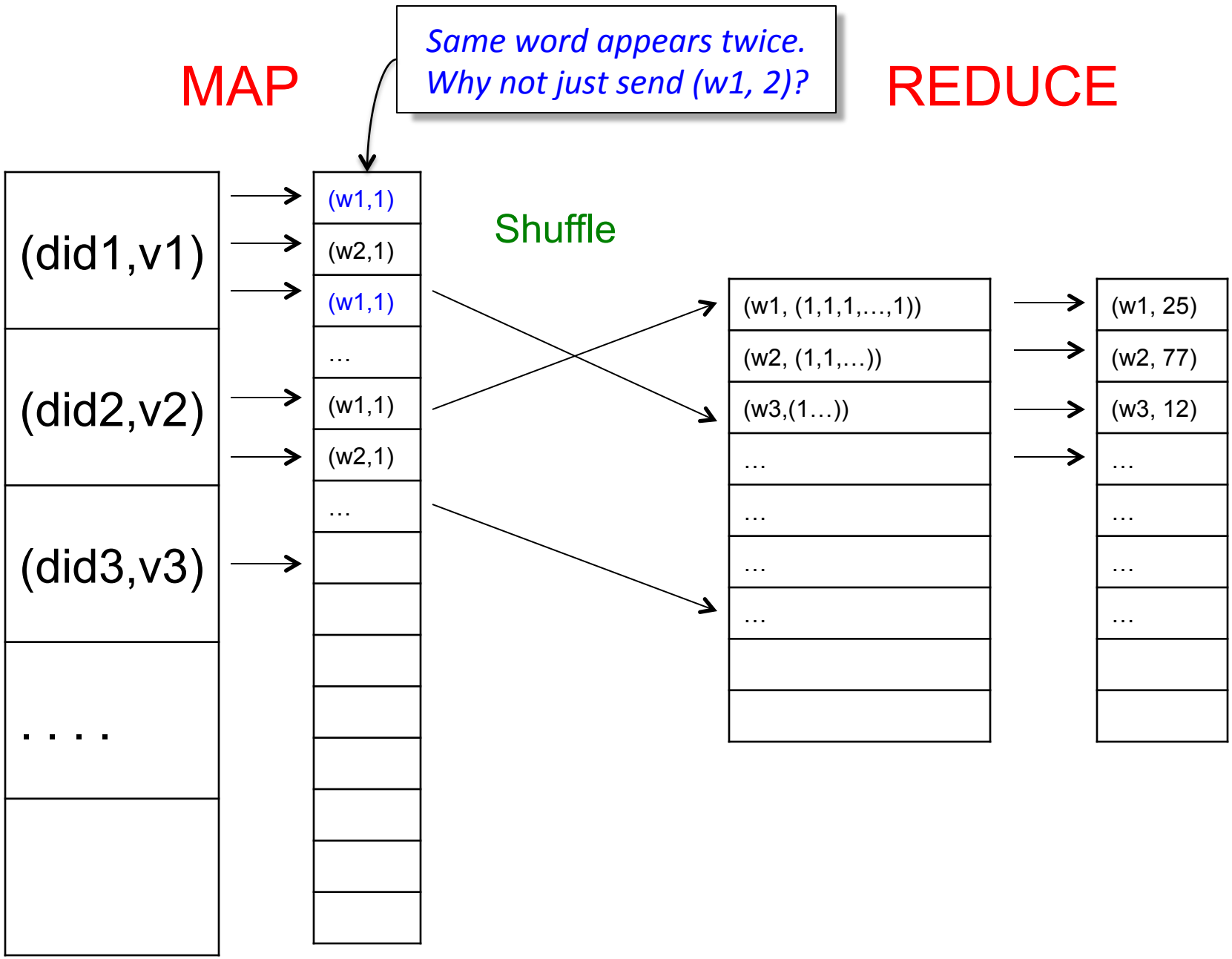
Distributed File System (DFS)

- For very large files: TBs, PBs
- Each file is partitioned into *chunks*, typically 64MB
- Each chunk is replicated several times (≥3), on different racks, for fault tolerance
- Implementations:
  - Google's DFS: GFS, proprietary
  - Hadoop's DFS: HDFS, open source

MR Phases

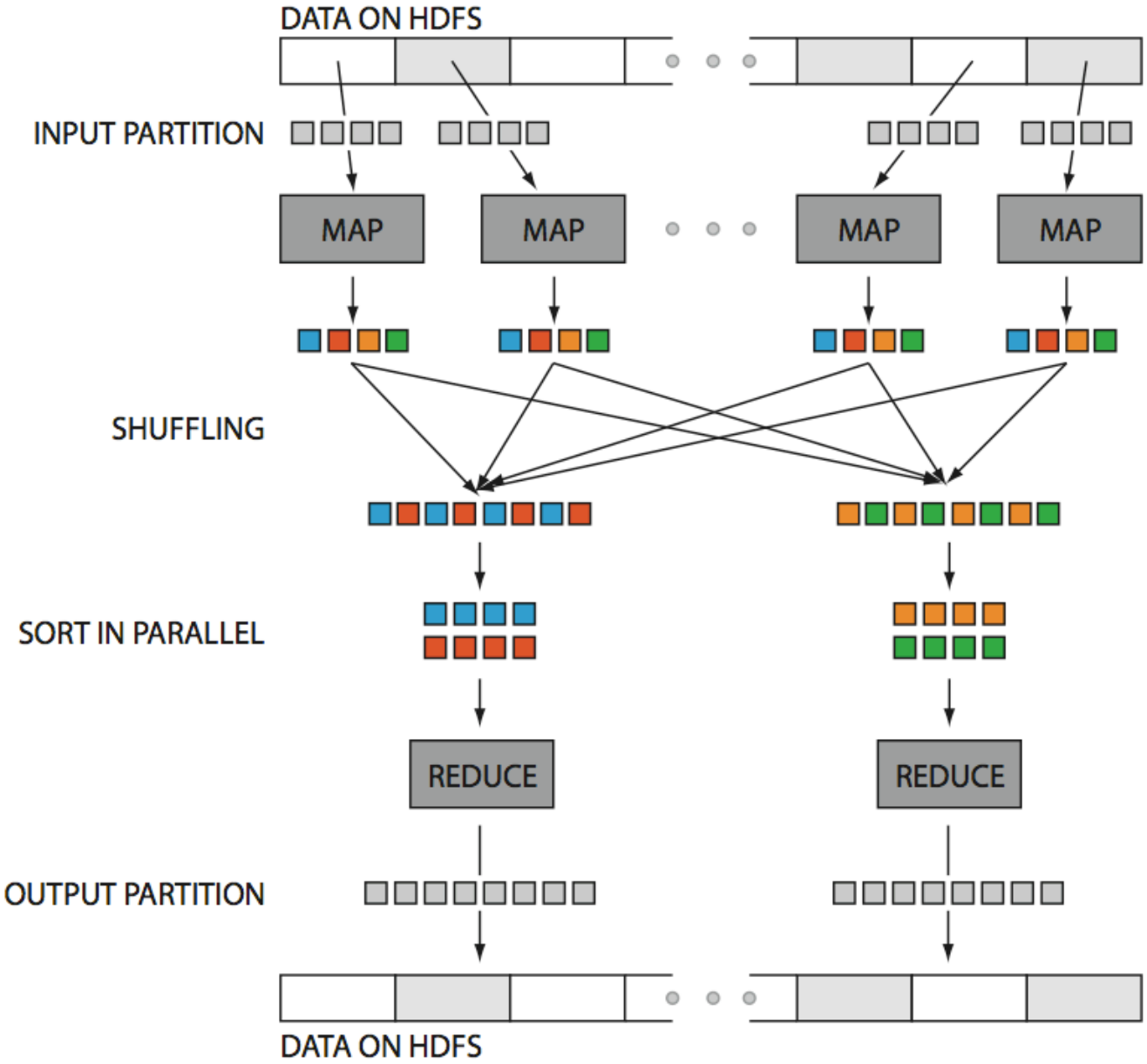
- Each Map and Reduce task has multiple phases:




MAP
Same word appears twice.
Why not just send (w1, 2)?
REDUCE
(did1,v1)
(did2,v2)
(did3,v3)
. . . .
(w1,1)
(w2,1)
(w1,1)
…
(w1,1)
(w2,1)
…
Shuffle
(w1, (1,1,1,…,1))
(w2, (1,1,…))
(w3,(1…))
…
…
…
…
(w1, 25)
(w2, 77)
(w3, 12)
…
…
…
…

Hadoop in One Slide

src: Huy Vo, NYU Poly

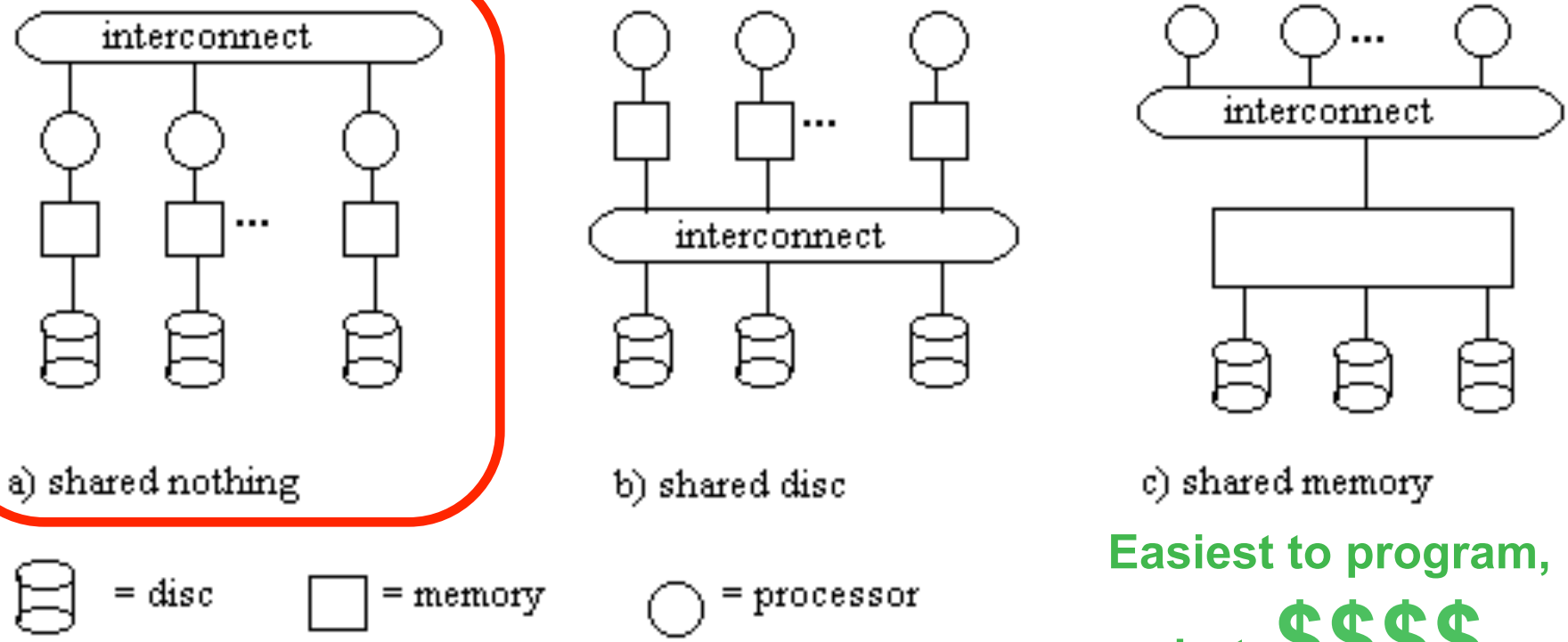


Distributed File System (DFS)

- For very large files: TBs, PBs
- Each file is partitioned into *chunks*, typically 64MB
- Each chunk is replicated several times (≥3), on different racks, for fault tolerance
- Implementations:
  - Google's DFS: GFS, proprietary
  - Hadoop's DFS: HDFS, open source

Taxonomy of Parallel Architectures

Scales to 1000s of computers

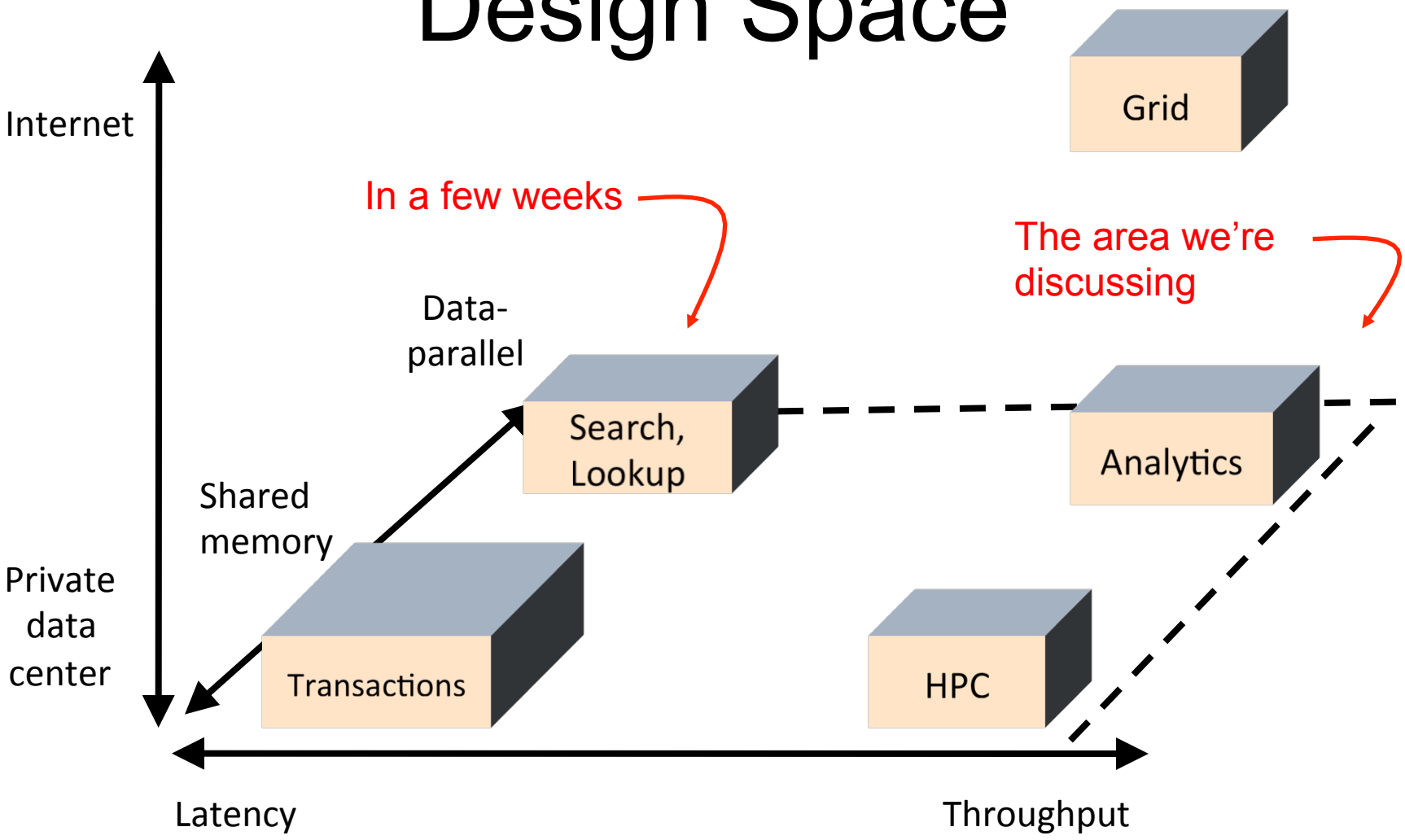


Easiest to program,

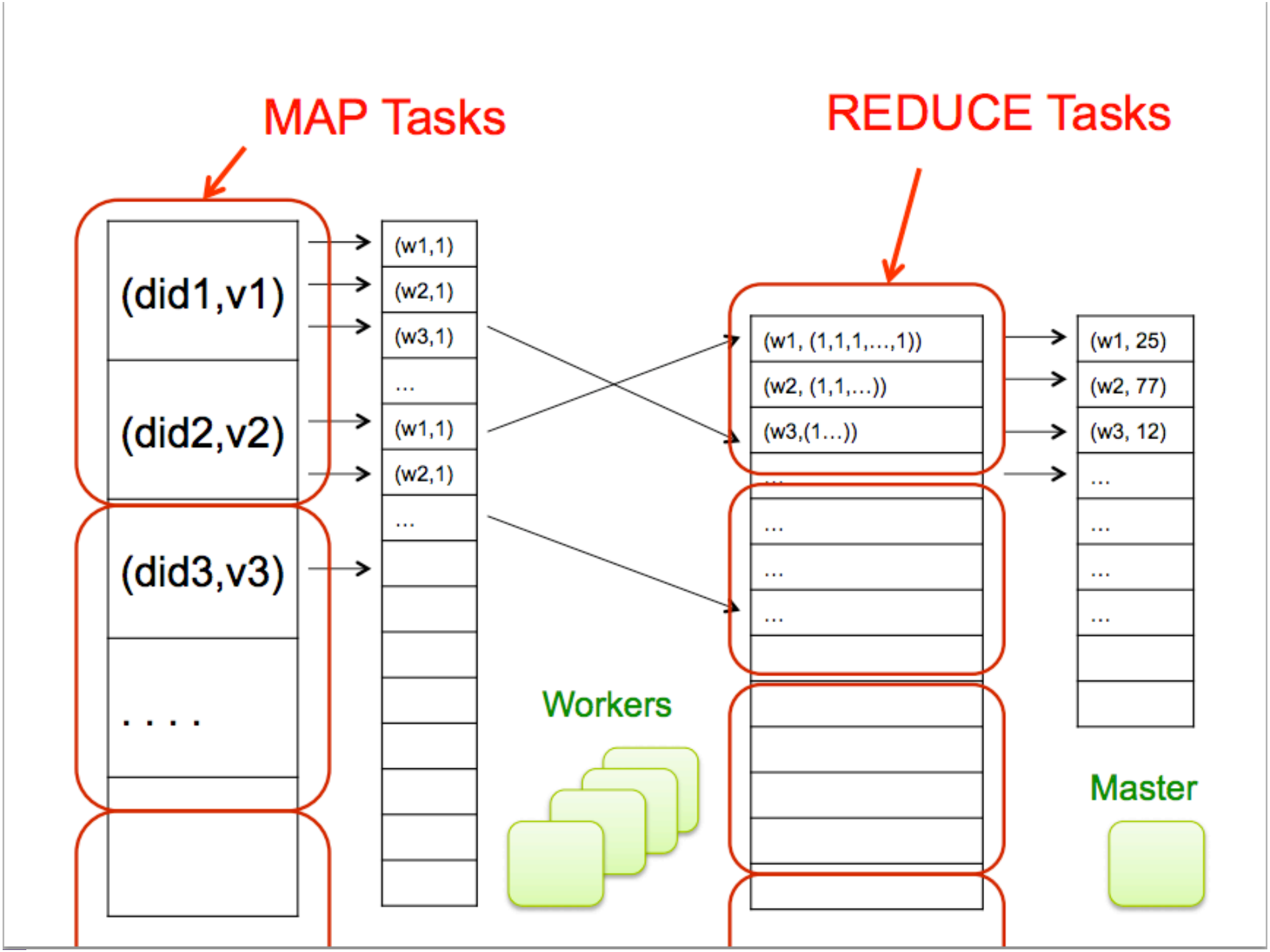
but $$$$

Fig. 3.1 Logical multi-processor database designs (diagram after [DEWI92])

Design Space



inspired by a slide by Michael Isard at Microsoft Research


MAP Tasks
REDUCE Tasks
(did1,v1)
(did2,v2)
(did3,v3)
. . . .
(w1,1)
(w2,1)
(w3,1)
…
(w1,1)
(w2,1)
…
(w1, (1,1,1,…,1))
(w2, (1,1,…))
(w3,(1…))
…
…
…
…
(w1, 25)
(w2, 77)
(w3, 12)
…
…
…
…
Workers
Master

Implementation

- There is one master node
- Master partitions input file into *M splits*, by key
- Master assigns *workers* (=servers) to the *M map tasks*, keeps track of their progress
- Workers write their output to local disk, partition into *R regions*
- Master assigns workers to the *R reduce tasks*
- Reduce workers read regions from the map workers' local disks

Large-Scale Data Processing

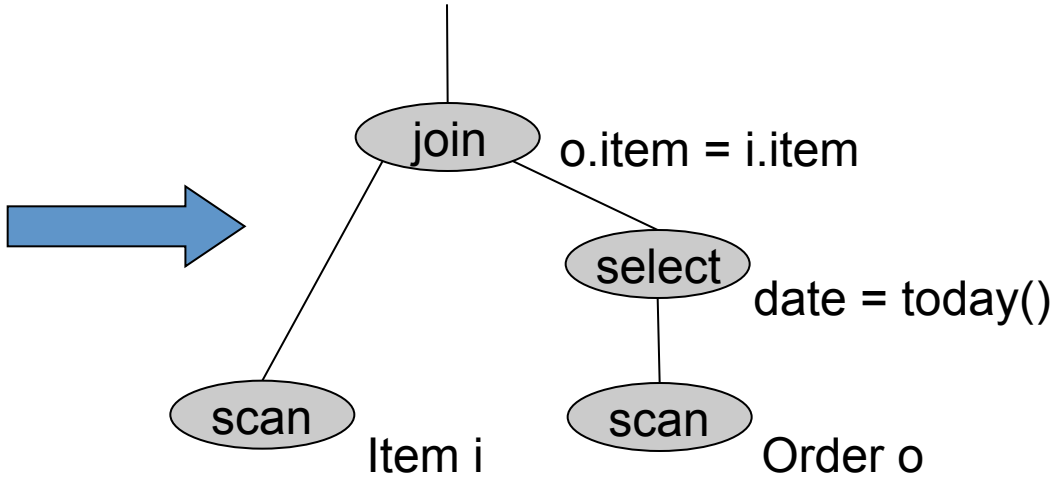
- Many tasks process big data, produce big data
- Want to use hundreds or thousands of CPUs
  - … but this needs to be easy
  - Parallel databases exist, but they are expensive, difficult to set up, and do not necessarily scale to hundreds of nodes.

- MapReduce is a *lightweight* framework, providing:
  - **Automatic parallelization and distribution**
  - **Fault-tolerance**
  - **I/O scheduling**
  - **Status and monitoring**

Key Idea: Declarative Languages

Find all orders from today, along with the items ordered

```
SELECT *
  FROM Order o, Item i
 WHERE o.item = i.item
   AND o.date = today()
```



Two notions of parallel query processing

- "Distributed Query"
  - Rewrite the query as a union of subqueries
  - Workers communicate through standard interfaces, so compatible with federated, heterogeneous, or distributed databases
- "Parallel Query"
  - Each operator is implemented with a parallel algorithm

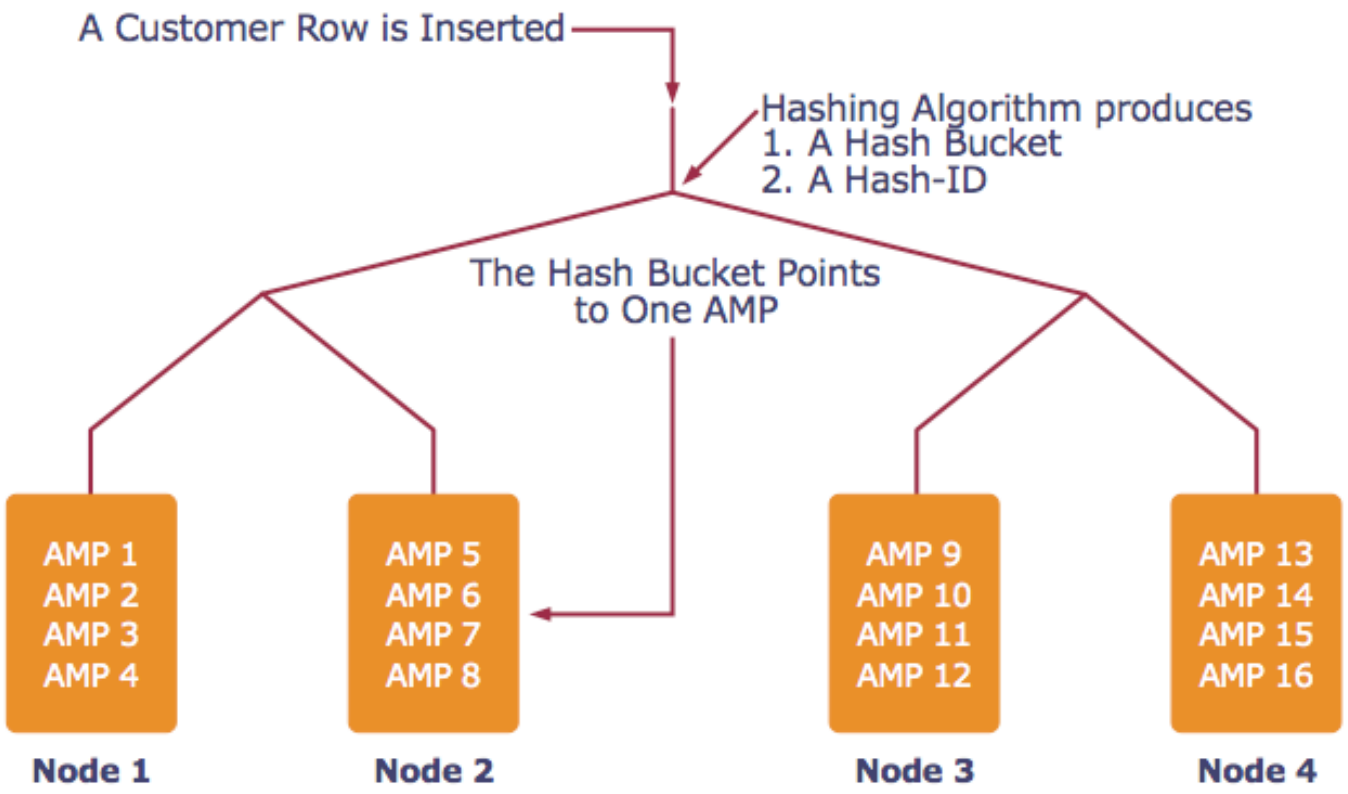
Distributed Query Example

```
CREATE VIEW Sales AS

SELECT * FROM JanSales
  UNION ALL
SELECT * FROM FebSales
  UNION ALL
SELECT * FROM MarSales
```

```
CREATE TABLE MarSales(
   OrderID       INT,
   CustomerID    INT       NOT NULL,
   OrderDate     DATETIME       NULL
      CHECK (DATEPART(mm, OrderDate) = 3),
  CONSTRAINT OrderIDMonth PRIMARY KEY(OrderID)
)
```

Parallel Query Example: Teradata

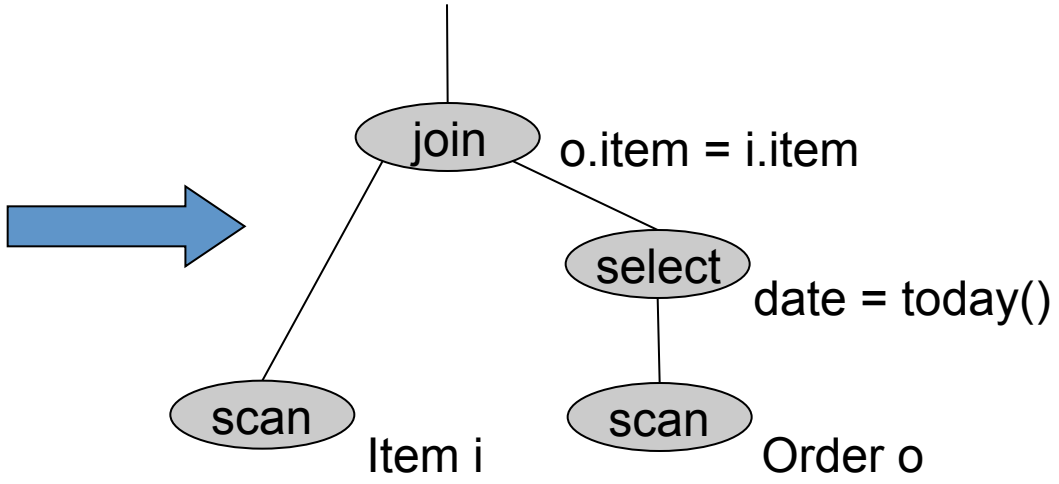


AMP = unit of parallelism

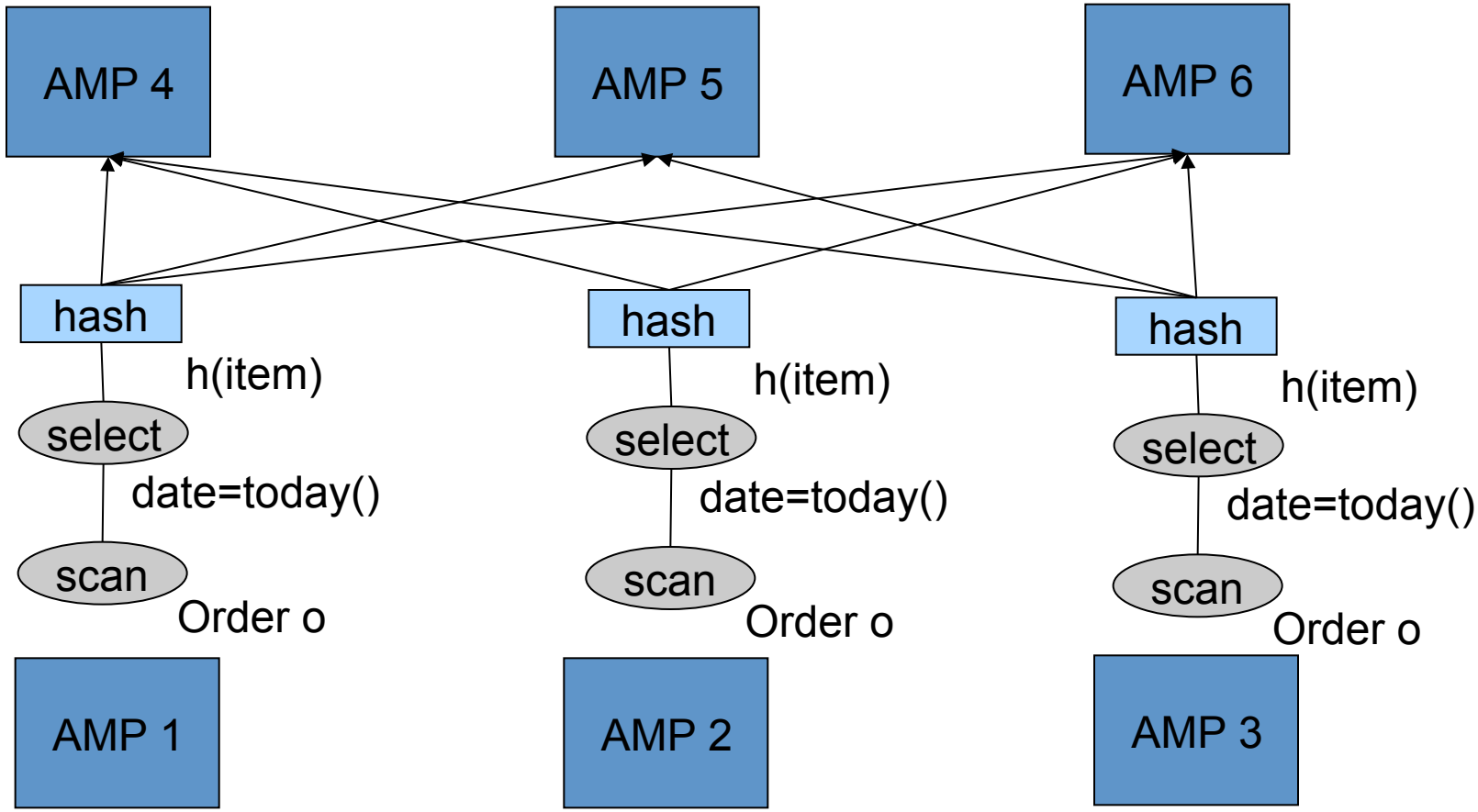
Example System: Teradata

Find all orders from today, along with the items ordered

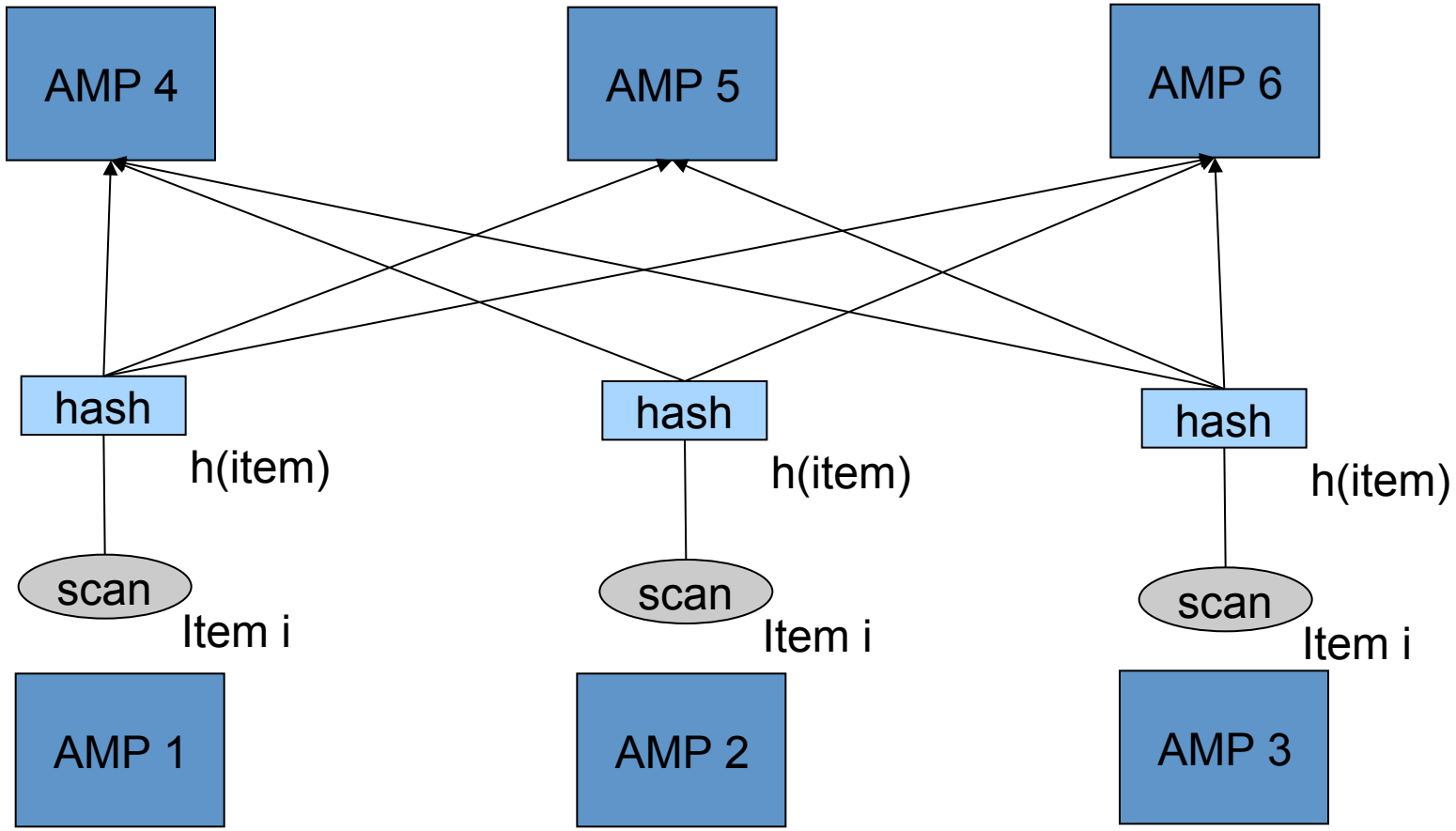
```
SELECT *
  FROM Orders o, Lines i
 WHERE o.item = i.item
   AND o.date = today()
```



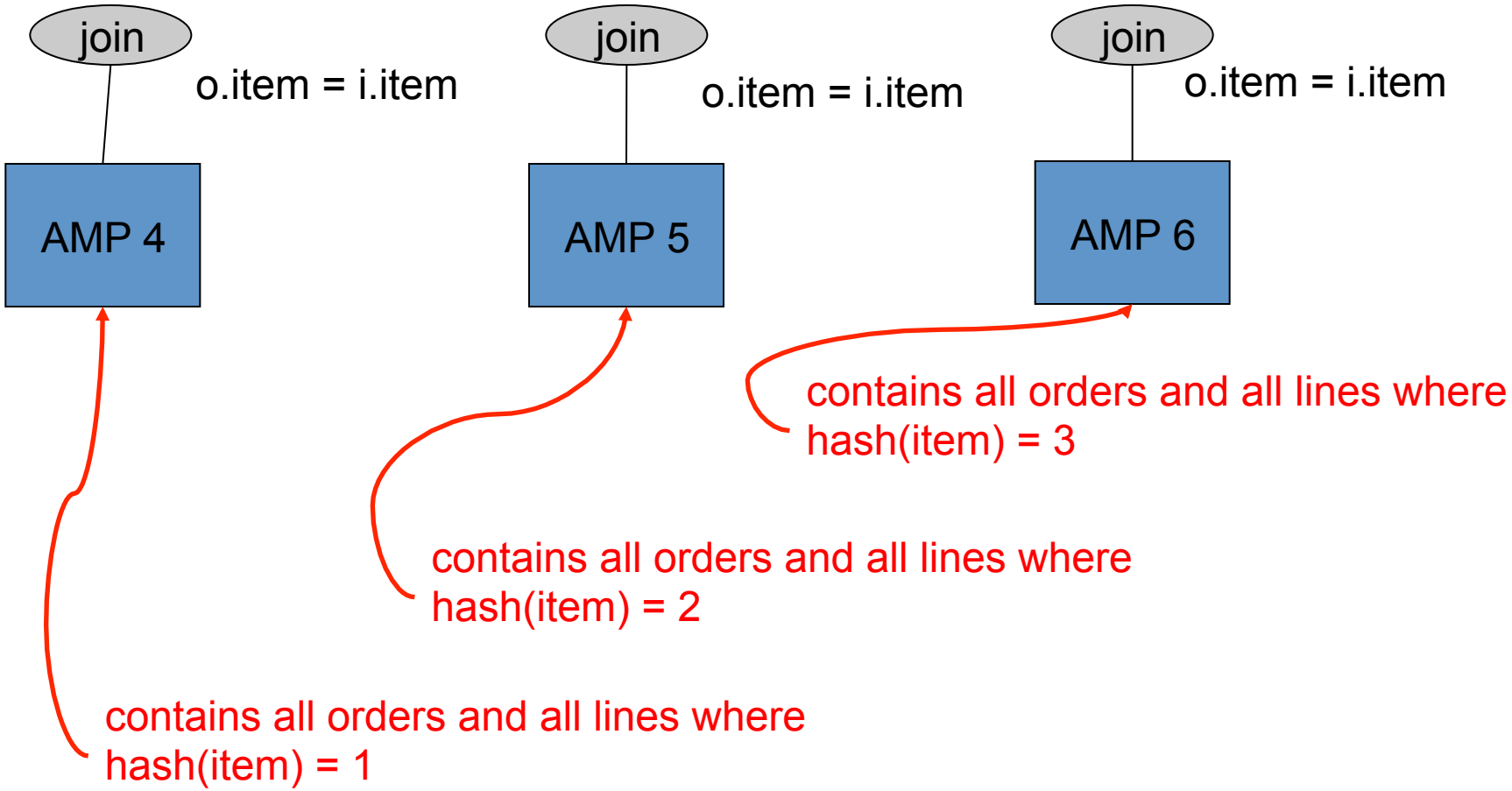
Example System: Teradata



Example System: Teradata



Example System: Teradata



MapReduce Contemporaries

- Dryad (Microsoft)
  - Relational Algebra
- Pig (Yahoo)
  - Near Relational Algebra over MapReduce
- HIVE (Facebook)
  - SQL over MapReduce
- Cascading
  - Relational Algebra
- Clustera
  - U of Wisconsin
- Hbase
  - Indexing on HDFS

MapReduce vs RDBMS

- RDBMS
  - Declarative query languages — HIVE, Pig, DryadLINQ
  - Schemas
  - Logical Data Independence — Hbase
  - Indexing — Pig, (Dryad, HIVE)
  - Algebraic Optimization
  - Caching/Materialized Views
  - *ACID/Transactions*

DryadLINQ, Pig, HIVE

- MapReduce
  - High Scalability
  - Fault-tolerance
  - "One-person deployment"

	Data Model	Prog. Model	Services
GPL	*	*	Typing (maybe)
Workflow	*	dataflow	typing, provenance, scheduling, caching, task parallelism, reuse
Relational Algebra	Relations	Select, Project, Join, Aggregate, …	optimization, physical data independence, data parallelism
MapReduce	[(key,value)]	Map, Reduce	massive data parallelism, fault tolerance
MS Dryad	IQueryable, IEnumerable	RA + Apply + Partitioning	typing, massive data parallelism, fault tolerance
MPI	Arrays/ Matrices	70+ ops	data parallelism, full control

Today's Reading, Pavlo 2009

MR VS. DATABASES

Pavlo et al. 2009

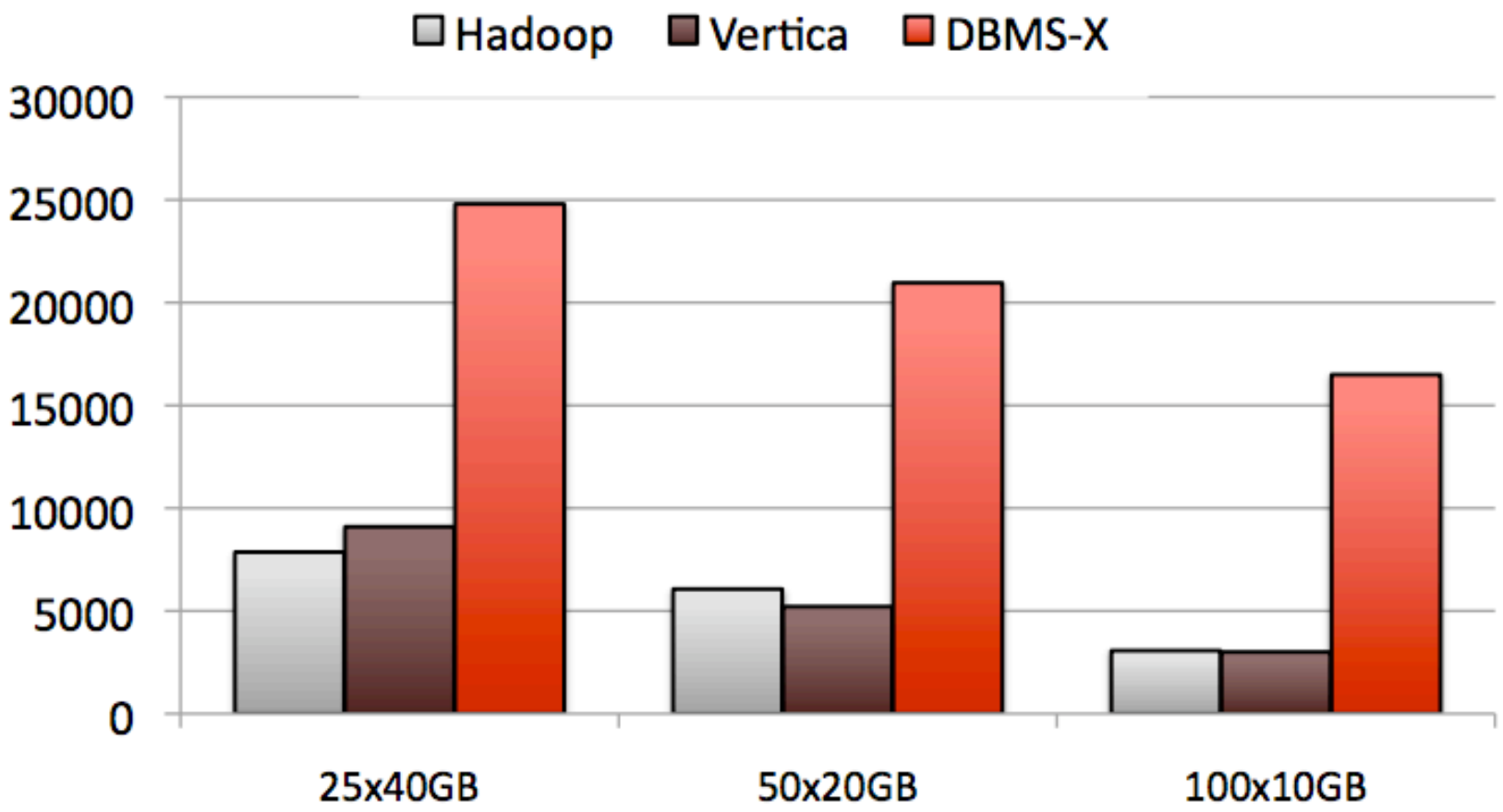
Hadoop vs. RDBMS

- Comparison of 3 systems
  - Hadoop
  - Vertica (a column-oriented database)
  - DBMS-X (a row-oriented database)
    - rhymes with “schmoracle”
- Qualitative
  - Programming model, ease of setup, features, etc.
- Quantitative
  - Data loading, different types of queries

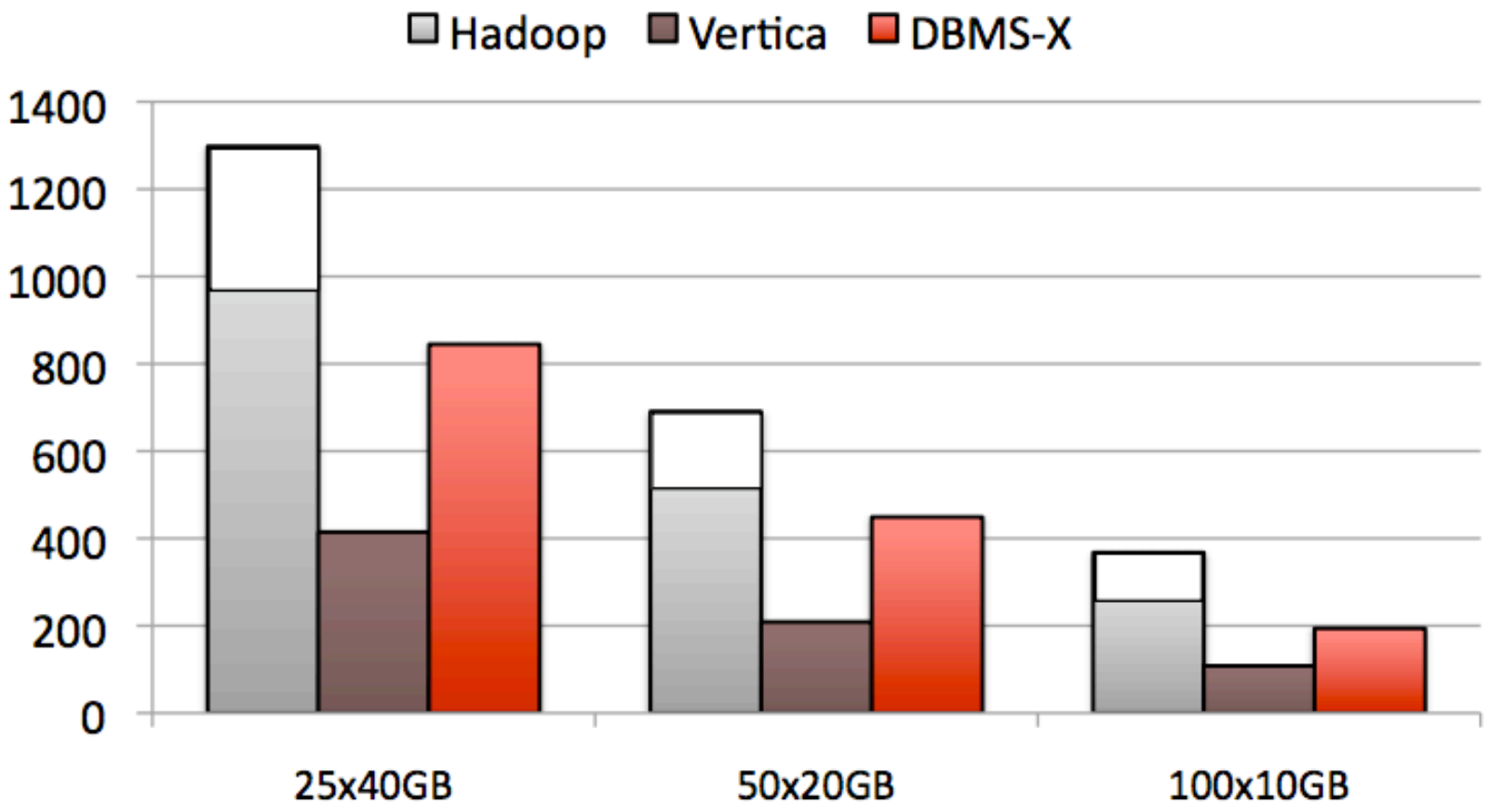
Grep Task

- Find 3-byte pattern in 100-byte record
  - *1 match per 10,000 records*
- Data set:
  - *10-byte unique key, 90-byte value*
  - *1TB spread across 25, 50, or 100 nodes*
  - *10 billion records*
- Original MR Paper (Dean et al. 2004)

Grep Task Loading Results

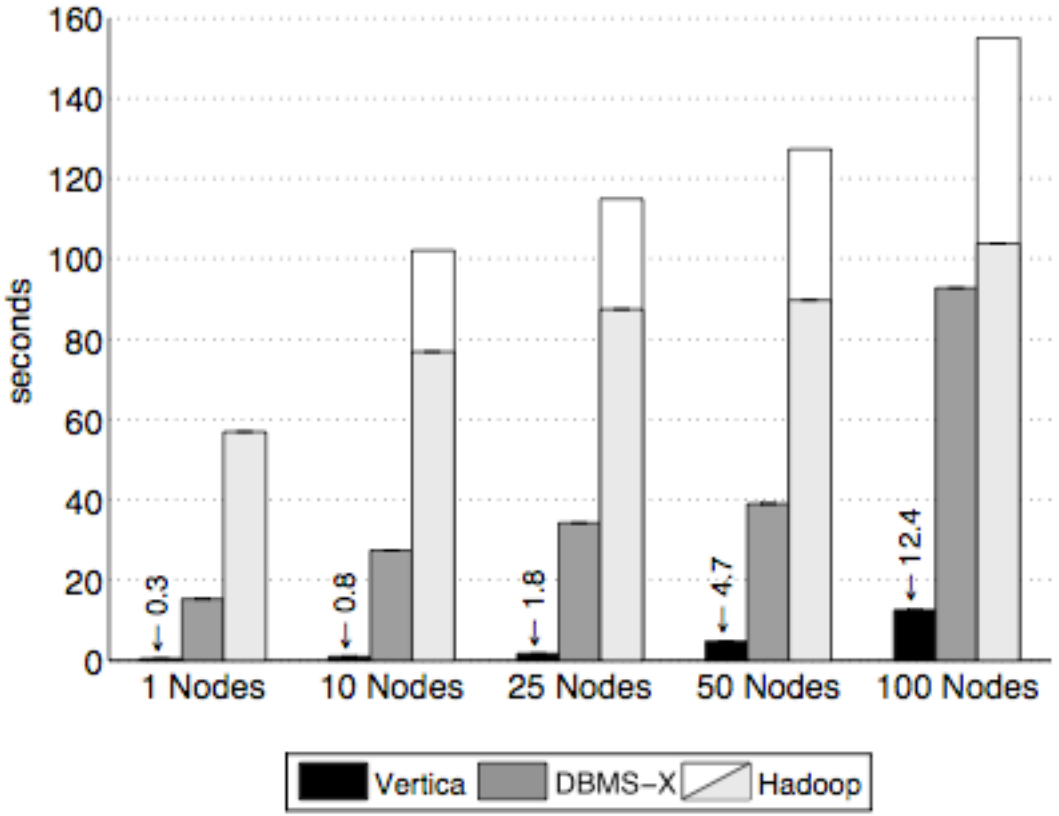


Grep Task Execution Results



Selection Task

```
SELECT pageURL, pageRank
FROM Rankings WHERE pageRank > X
```



1 GB / node

Analytical Tasks

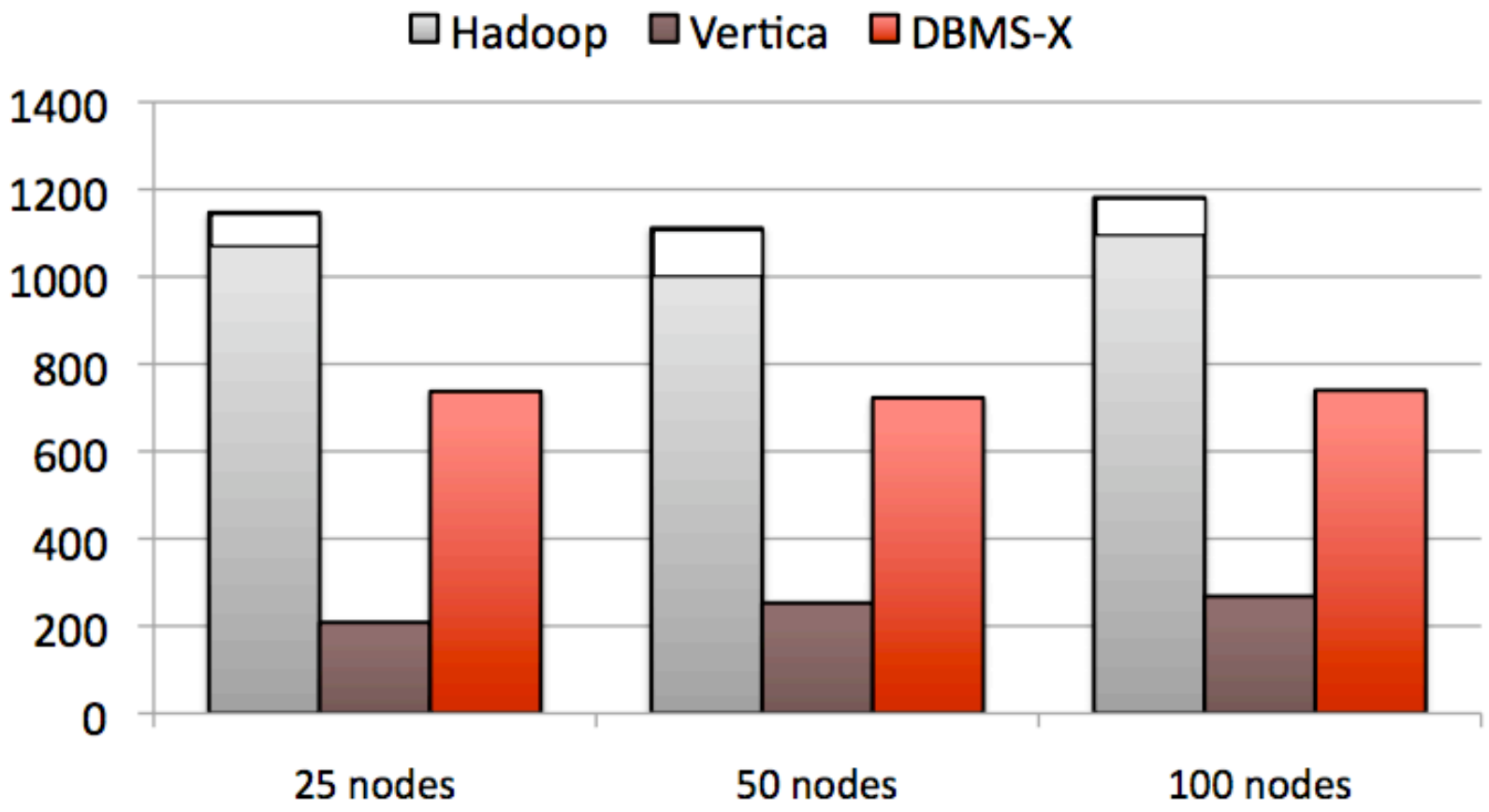
- Simple web processing schema
- Data set:
  - *600k HTML Documents (6GB/node)*
  - *155 million UserVisit records (20GB/node)*
  - *18 million Rankings records (1GB/node)*

Aggregate Task

- Simple query to find adRevenue by IP prefix

```
SELECT SUBSTR(sourceIP, 1, 7),
       SUM(adRevenue)
  FROM userVistits
 GROUP BY SUBSTR(sourceIP, 1, 7)
```

Aggregate Task Results



Join Task

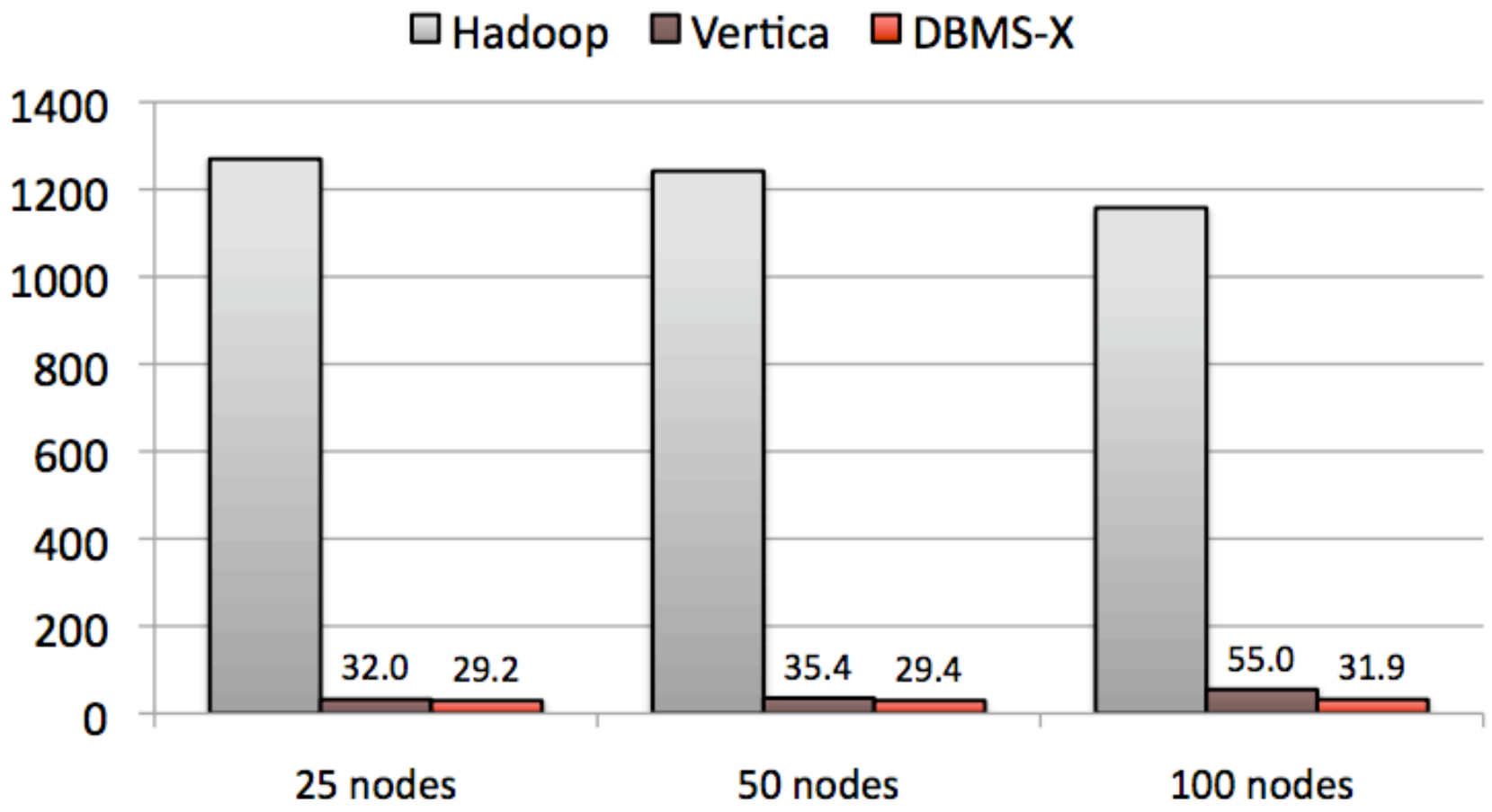
- **Find the sourceIP that generated the most adRevenue along with its average pageRank.**
- **Implementations:**
  - ***DBMSs – Complex SQL using temporary table.***
  - ***MapReduce – Three separate MR programs.***

Join Task

```
SELECT INTO TempsourceIP,
            AVG(pageRank)as avgPageRank,
            SUM(adRevenue)as totalRevenue
FROM RankingsAS R
   , UserVisitsAS UV
WHERE R.pageURL = UV.destURL
AND UV.visitDate
  BETWEEN '2000-01-15'
  AND '2000-01-22'
GROUP BY UV.sourceIP;

SELECT sourceIP,
       totalRevenue,
       avgPageRank
FROM Temp
ORDER BY totalRevenueDESC
LIMIT 1;
```

Join Task Results



Problems with this analysis?

- Other ways to avoid sequential scans?
- Fault-tolerance in large clusters?
- Tasks that cannot be expressed as queries?

Google's Response: Cluster Size

- **Largest known database installations:**
  - ***Greenplum – 96 nodes – 4.5 PB (eBay) [1]***
  - ***Teradata – 72 nodes – 2+ PB (eBay) [1]***
- **Largest known MR installations:**
  - ***Hadoop – 3658 nodes – 1 PB (Yahoo) [2]***
  - ***Hive – 600+ nodes – 2.5 PB (Facebook) [3]***

[1] eBay's two enormous data warehouses – April 30th, 2009
http://www.dbms2.com/2009/04/30/ebays-two-enormous-data-warehouses/
[2] Hadoop Sorts a Petabyte in 16.25 Hours and a Terabyte in 62 Seconds – May 11th, 2009
http://developer.yahoo.net/blogs/hadoop/2009/05/hadoop_sorts_a_petabyte_in_162.html
[3] Hive - A Petabyte Scale Data Warehouse using Hadoop – June 10th, 2009
http://www.facebook.com/note.php?note_id=89508453919

Concluding Remarks

- What can *MapReduce* learn from *Databases*?
  - *Declarative languages are a good thing.*
  - *Schemas are important.*
- What can *Databases* learn from *MapReduce*?
  - *Query fault-tolerance.*
  - *Support for in situ data.*
  - *Embrace open-source.*

Other Benchmarked Systems

- **HadoopDB (Abadi '09 - Yale)**
  - ***Replaced Hadoop filesystem with Postgres.***
  - ***Makes JDBC calls inside of MR functions.***
- **Hive (Thusoo '09 - Facebook)**
  - ***Data warehouse interface on top of Hadoop.***
  - ***Converts SQL-like language to MR programs.***