General Purpose Computing Systems I : Hadoop and MapReduce

Shiow-yang Wu (吳秀陽)

CSIE, NDHU, Taiwan, ROC

Lecture material is mostly home-grown, partly taken with permission and courtesy from Professor Shih-Wei Liao of NTU.

Outline

- What is Hadoop? Why so popular? Still now?
- What is MapReduce? Why MapReduce? What is it used for?
- MapReduce concepts, models and examples
- Hadoop cluster for MapReduce
- Execution details and internals
- Problems with Hadoop and MapReduce
- Current status

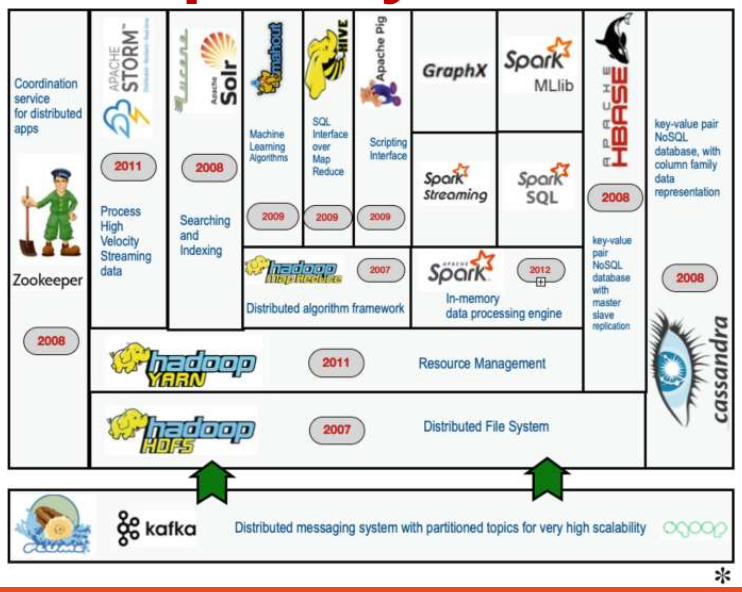
Outline (cont.)

- Problem solving and algorithm design with MapReduce
- Some MapReduce algorithms
- Data mining with MapReduce
- Hadoop and MapReduce practice (Assignment)

What is Hadoop?

- Apache Hadoop is an 100% open source framework for "reliable, scalable, distributed computing" on large volume of data across **clusters** of commodity hardware.
- First released on 2006 based on Google's MapReduce(more about this later).
- Over years of development into Hadoop ecosystem, the framework has become one of the most prominent and used open-source tool in big data era.
- Latest release: 3.3.6 (2023 Jun 23) (https://hadoop.apache.org/)

Hadoop Ecosystem



Hadoop Ecosystem

(https://www.cloudduggu.com/hadoop/ecosystem/)

Examples of Tools on Hadoop

- Lots of useful tools are supported on Hadoop

Tool	Description	First Release	Latest Update
YARN	Resource Manager & Scheduler	2006	2023-06-18
Hbase	no-SQL database	2008	2023-06-13
Hive	Data Warehouse and SQL abstraction	2010	2023-08-14
Sqoop	RDMBS ingestion pipeline	2009	2019-01-18
Spark	Data processing framework and compute engine	2014	2023-09-13
Tez	Execution frameworks for DAGS on Hive or Pig	2014	2022-06-15

- Rumors of "Hadoop is dying/dead!" never stops.
- From the dates of latest update, Hadoop is alive and kicking !! (We will keep watching.)

Why Hadoop?

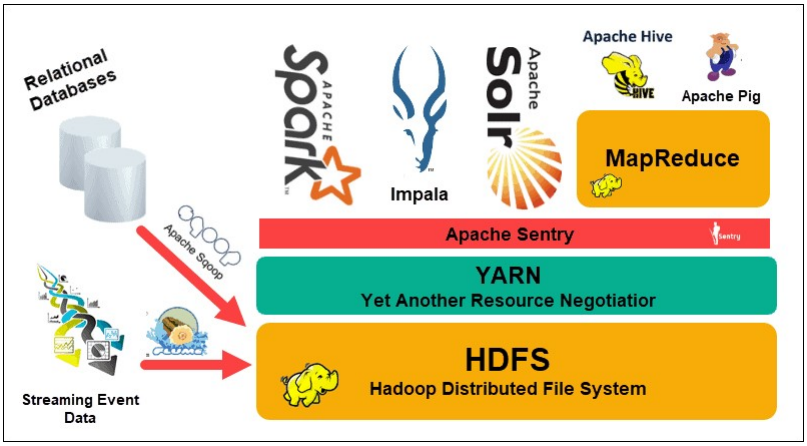
- Flexible and versatile
- Scalable and cost effective
- More efficient data economy
- Rich and robust Ecosystem
- Hadoop is getting more "Real-Time"!
- New technologies still active on Hadoop
- Remain one of the top open-source big data tools.
- However, the future of Hadoop is cloudy.

Hadoop Alternatives

- The latest stable release of Hadoop is 3.3.6 (2023-06-23) of the 3.3 line.
- There are problems with Hadoop. (later)
- Newer feature-packed systems w/o those problems become popular Hadoop alternatives.
- Apache Spark is a good example. (later)
- Hadoop is no longer dominate but still popular.
- Hadoop is also good for learning purpose.

Hadoop Architecture

- HDFS, YARN, and MapReduce are at the heart of the Hadoop ecosystem.



What is MapReduce?

- Data-parallel programming model for clusters of commodity machines
  - Designed for scalability and fault-tolerance
- Pioneered by Google
  - Processes 20 PB of data per day (at that time)
- Popularized by open-source Hadoop project
  - Used by Yahoo!, Facebook, Amazon, …
- Google stop using MapReduce in favor of newer tools (Dataflow, Apache Beam, …)
- MR keeps involving and still popular

MapReduce Usage Examples

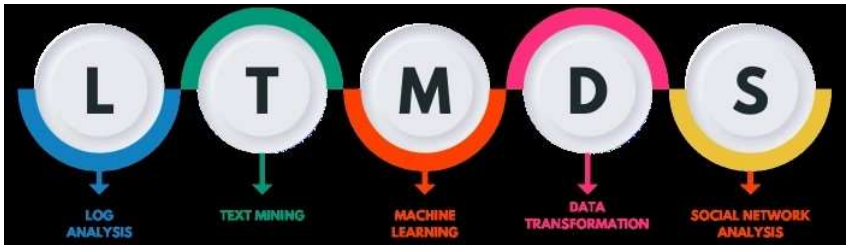
- At Google:
  - Index building for Google Search
  - Article clustering for Google News
  - Statistical machine translation
- At Yahoo!:
  - Index building for Yahoo! Search
  - Spam detection for Yahoo! Mail
- At Facebook:
  - Data mining
  - Ad optimization
  - Spam detection

MapReduce Usage Examples

- In research:
  - Analyzing Wikipedia conflicts (PARC)
  - Natural language processing (CMU)
  - Bioinformatics (Maryland)
  - Particle physics (Nebraska)
  - Ocean climate simulation (Washington)
  - Sequential pattern mining (NDHU)
  - <Your applications here>

Power of MapReduce

- MR can be considered as a parallel computing model which can be used in many different areas.
- It has been shown that any problem in *NC* (problems efficiently solvable on a parallel computer) can be efficiently solved with MR.



Hadoop/MR History

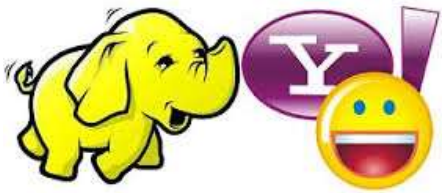
- The foundation stone: The Google File System by Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung in 2003. (*19th ACM Symposium on Operating Systems Principles*)
- The paper that started everything – MapReduce: Simplified Data Processing on Large Clusters by Jeffrey Dean and Sanjay Ghemawat in 2004. (*6th Symposium on Operating System Design and Implementation*)
- Reading 2 papers above is strongly recommended!

Hadoop/MR History

- Shortly after the MapReduce paper, open source pioneers Doug Cutting and Mike Cafarella started working on a MapReduce implementation to solve the scalability problem of Nutch (an open source search engine)
- Over the course of a few months, Cutting and Cafarella built up the underlying file systems and processing framework that would become Hadoop (in Java)

Hadoop/MR History

- In 2006, Cutting went to work for Yahoo.
- They spun out the storage and processing parts of Nutch to form Hadoop (named after Cutting's son's stuffed elephant).
- Over time and heavy investment by Yahoo!, Hadoop eventually became a top-level Apache Foundation project.



Hadoop/MR Today

- Over the years, numerous independent people and organizations contribute to Hadoop.
- Every new release adds functionality and boosts performance.
- Several other open source projects have been built with Hadoop at their core, and this list is continually growing.
- Some of the more popular ones: Pig(programming tool), Hive(warehousing), HBase(NoSQL DB), Mahout(machine learning), and ZooKeeper(distributed systems and services).

Why Hadoop/MapReduce?

- **Problem**: Lots of data!
- Example: Word frequencies in Web pages
- This is how the early Internet search engines (Archie, AltaVista) was done.
- Search results are ordered based on keyword frequencies.
- The new engine from Larry and Sergey's project at Stanford revolutionized the search industry. (Google)
- (https://blog.reputationx.com/anatomy-of-search-results)

Word Frequencies in Pages

- Given 130 trillion web pages x 1KB/page = 130PB
- If one computer can read 750 MB/sec from disk
  - 5+ years to read the web
  - 13K hard drives(10TB HDD) to store the web
- Even more: To do something with the data
  - Compute the word frequencies for each word in each website

Basic Solution: Spread the work over many machines

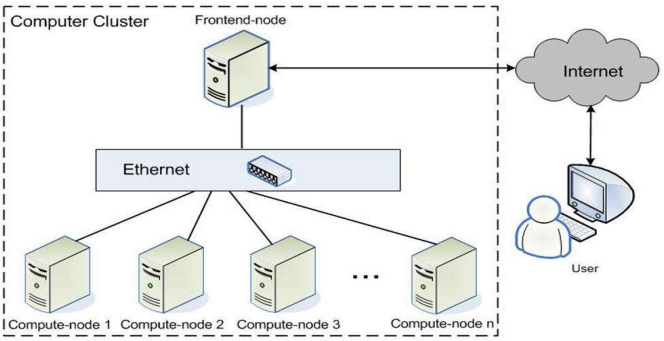
- Same problem with 10,000 machines: 4+ hours
- New problems: Extra programming works
  - communication and coordination
  - recovering from machine failure
  - status reporting
  - debugging
  - optimization
  - locality
- Those works repeat for every problem you want to solve

Hadoop/MR Design Goals

1. **Scalability to large data volumes:**
   - Scan 100 TB on 1 node @ 50 MB/s = 24 days
   - Scan on 1000-node cluster = 35 minutes
   - => 1000's of machines, 10,000's of disks
2. **Cost-efficiency:**
   - Commodity machines (cheap, but unreliable)
   - Commodity network
   - Automatic fault-tolerance (fewer administrators)
   - Easy to use (less programming)

Computing Clusters

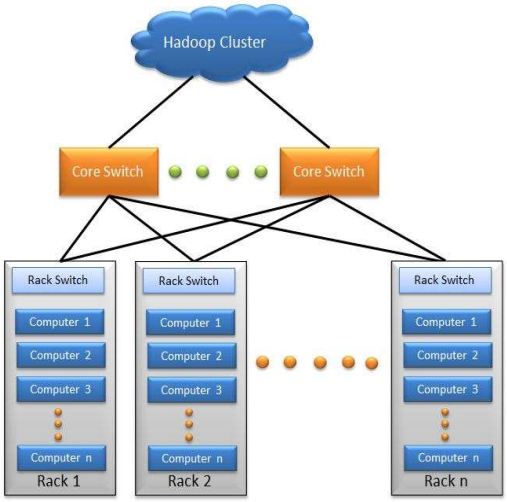
- Many racks of computers, thousands of machines per cluster
- Limited bisection bandwidth between racks



Hadoop Cluster

Typical Hadoop Cluster

- 30-40 nodes/rack, 1000-4000 nodes in cluster
- 1 Gbps within rack, 10 Gbps across racks
- H/W specs: depends on the deployment mode. (next slide)



Hardware	Sandbox Deployment	Basic or Standard Deployment	Advanced Deployment
CPU speed	2 - 2.5 GHz	2 - 2.5 GHz	2.5 - 3.5 GHz
Logical or virtual CPU cores	16	24 - 32	48
Total system memory	16 GB	64 GB	128 GB
Local disk space for yarn.nodemanager.local-dirs[1]	256 GB	500 GB	2.4 TB
DFS block size	128 MB	256 MB	256 MB
HDFS replication factor	3	3	3
Disk capacity	32 GB	256 GB - 1 TB	1.2 TB
Total number of disks for HDFS	2	8	12
Total HDFS capacity per node	64 GB	2 - 8 TB	At least 14 TB
Number of nodes	2 +	4 - 10+	12 +
Total HDFS capacity on the cluster	128 GB	8 - 80 TB	144 TB
Actual HDFS capacity (with replication)	43 GB	2.66 TB	57.6 TB
/tmp mount point	20 GB	20 GB	30 GB
Installation disk space requirement	12 GB	12 GB	12 GB
Network bandwidth (Ethernet card)	1 Gbps	2 Gbps (bonded channel)	10 Gbps (Ethernet card)

Implications of Computing Environment

- Single-thread performance doesn't matter
  - **Large problems** and **total throughput/$** are more important than peak performance
- Stuff Breaks
  - More nodes imply higher probability of breaking down
- "Ultra-reliable" hardware doesn't really help
  - At large scales, super-fancy reliable hardware still fails, albeit less often
    - software still needs to be fault-tolerant
    - commodity machines without fancy hardware give better perf/$

Challenges & Solutions

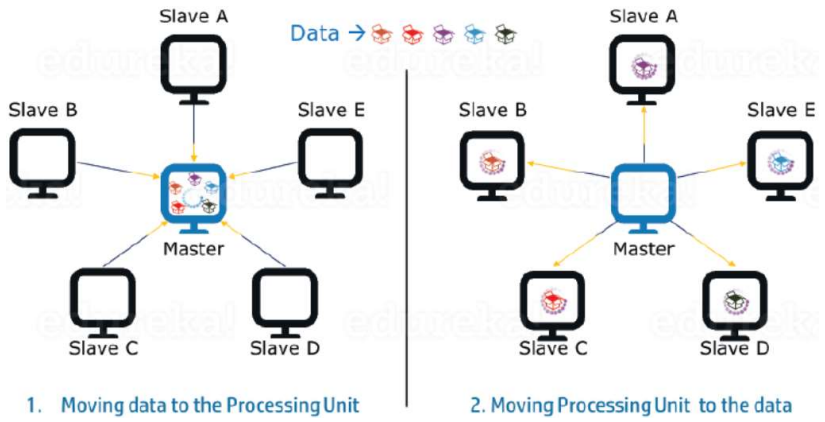
1. **Cheap nodes fail, especially if you have many**
   - Mean time between failures for 1 node = 3 years
   - Mean time between failures for 1000 nodes = 1 day
   - Solution: Build fault-tolerance into system
2. **Commodity network = low bandwidth**
   - Solution: Push computation to the data
3. **Programming distributed systems is hard**
   - Solution: Data-parallel programming model: users write "map" & "reduce" functions, system distributes work and handles faults

MapReduce

- A simple **programming model** that applies to many large-scale computing problems
- Hide messy details of distributed programs behind MapReduce runtime library:
  - Automatic parallelization
  - Load balancing
  - Network and disk transfer optimization
  - Handling of machine failures
  - Robustness
  - Improvements to core library

Traditional vs MapReduce

- MapReduce is basically a Divide&Conquer approach with a different idea of data/computation movement.



MapReduce Basics

- Semantics borrowed from ***function programming languages***
- FP language are usually **stateless,** which is very good for parallelism
  - No need to worry about synchronization
- Even not using MapReduce, many programming languages like Ruby and Python provides such semantics.
  - Simplicity
  - Chance for implicit optimization
  - Easily parallelizable

Functional Abstractions Hide Parallelism

- The ideas of functions, mapping and reducing are from functional programming languages (eg. Lisp)
- Map()
  - In FP: [1,2,3,4] - (*2) -> [2,4,6,8]
  - Process a key/value pair to generate intermediate key/value pairs
- Reduce()
  - In FP: [1,2,3,4] - (sum) -> 10
  - Merge all intermediate values associated with the same key
- Both Map and Reduce are easy to parallelize

MapReduce in Functions

- Data type: *key-value records*
- Map function:

$$(K_{in}, V_{in}) \rightarrow list(K_{inter}, V_{inter})$$

- Reduce function:

$$(K_{inter}, list(V_{inter})) \rightarrow list(K_{out}, V_{out})$$

Programming Model

1. Read a lot of data
2. **Map**: extract something you care about from each record
3. Shuffle and Sort
4. **Reduce**: aggregate, summarize, filter, or transform
5. Write the results

Outline stays the same,
map and reduce change to fit the problem

Programming Model: More Specifically

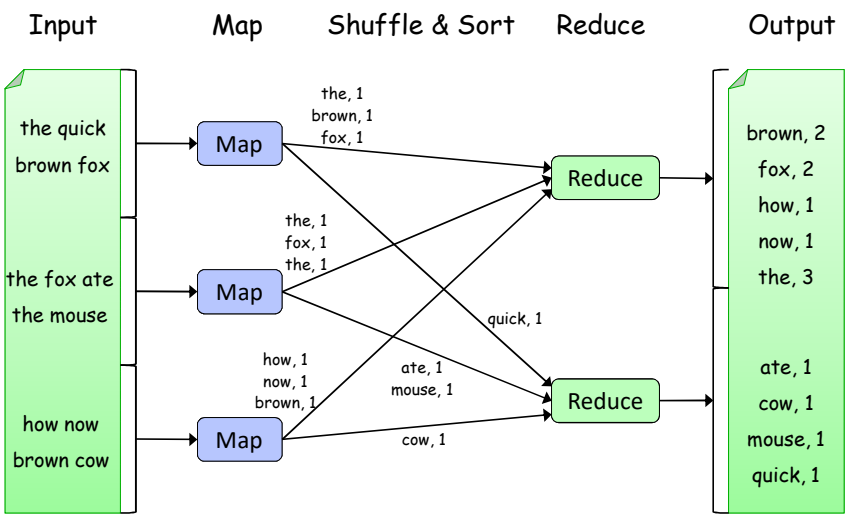
- Programmer specifies two primary methods:
  - `map(k, v) -> <k', v'>*`
  - `reduce(k', <v'>*) -> <k'', v''>*`
- All v' with same k' are reduced together ***in order***
- Can also specify:
  - partition(k' , total partitions) -> partition for k'
    - often a simple hash of the key
    - allows reduce operations for different k' to be parallelized

The Word Count Example

```
def mapper(line):
    foreach word in line.split():
        output(word, 1)

def reducer(key, values):
    output(key, sum(values))
```

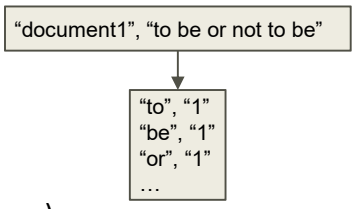
Word Count Execution

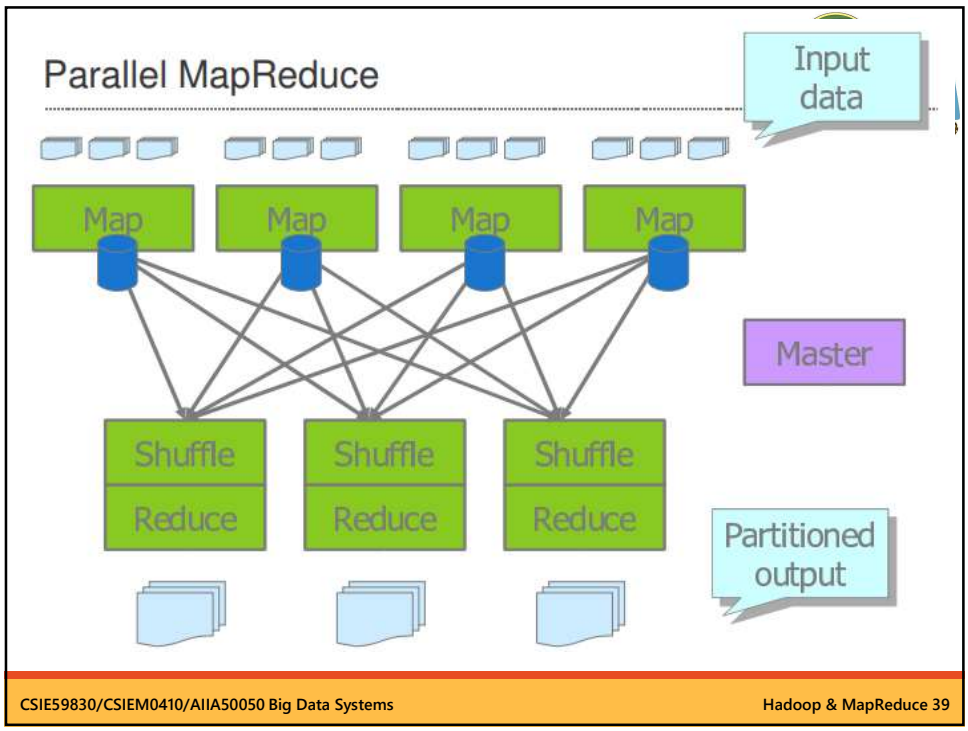


Web Pages Example

Example: Word Frequencies in Web Pages

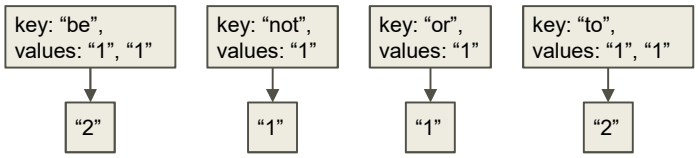
- Input is files with one document per record
- Specify a map function that takes a key/value pair
  - key: document URL
  - value: document contents
- Output of map function is (potentially many) key/value pairs.
- In our case, output (word, "1") once per word in the document



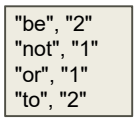


Web Pages Example (cont.)

- MapReduce library gathers together all pairs with the same key (shuffle/sort)
- The reduce function combines the values for a key. In our case, compute the sum



- Output of reduce paired with key and saved



Pseudo-Code

```
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 key, Iterator intermediate_values):
// key: a word, same for input and output
// intermediate_values: a list of counts
   int result = 0;
   for each v in intermediate_values:
     result += ParseInt(v);
   emit(asString(result));
```


How MapReduce Works

- User to do list:
  - indicate:
    - Input/output files
    - **M**: number of map tasks
    - **R**: number of reduce tasks
    - **W**: number of machines
  - Write *map* and *reduce* functions
  - Submit the job
- This requires no knowledge of parallel/distributed systems!!!
- What about everything else?

Data Distribution

- Input files are split into **M** pieces on distributed file system
  - Typically ~ 64 MB blocks
- Intermediate files created from *map* tasks are written to local disk
- Output files are written to distributed file system

Assigning Tasks

- Many copies of user program are started
- Tries to utilize data localization by running *map* tasks on machines with data
- One instance becomes the **Master**
- Master finds idle machines and assigns them tasks

Execution (map)

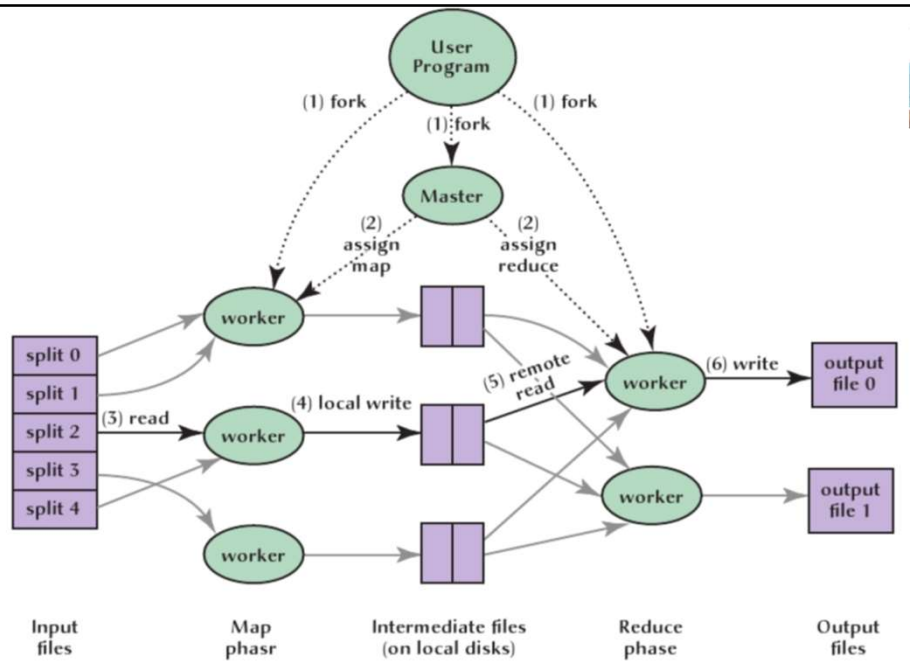
- *Map* workers read in contents of corresponding input partition
- Perform user-defined *map* computation to create intermediate <key, value> pairs
- Periodically buffered output pairs written to local disk
  - Partitioned into **R** regions by a partitioning function

Partition Function

- Example partition function: hash(key) mod **R**
- Why do we need this?
- Example Scenario:
  - Want to do word counting on 10 documents
  - 5 *map* tasks, 2 *reduce* tasks

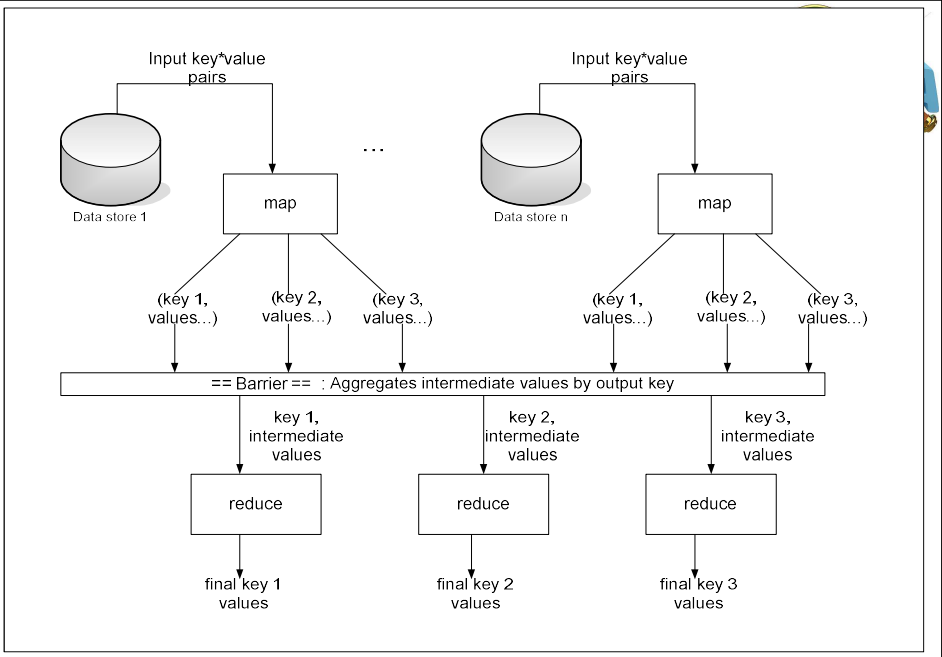
Execution (reduce)

- Reduce workers iterate over ordered intermediate data
  - Each unique key encountered – values are passed to user's reduce function
  - eg. <key, [value1, value2,..., valueN]>
- Output of user's *reduce* function is written to output file on distributed file system
- When all tasks have completed, master wakes up user program



Observations

- No *reduce* can begin until *map* is complete
- Tasks scheduled based on location of data
- If *map* worker fails any time before *reduce* finishes, task must be completely rerun
- Master must communicate locations of intermediate files
- MapReduce library does most of the hard work for us!



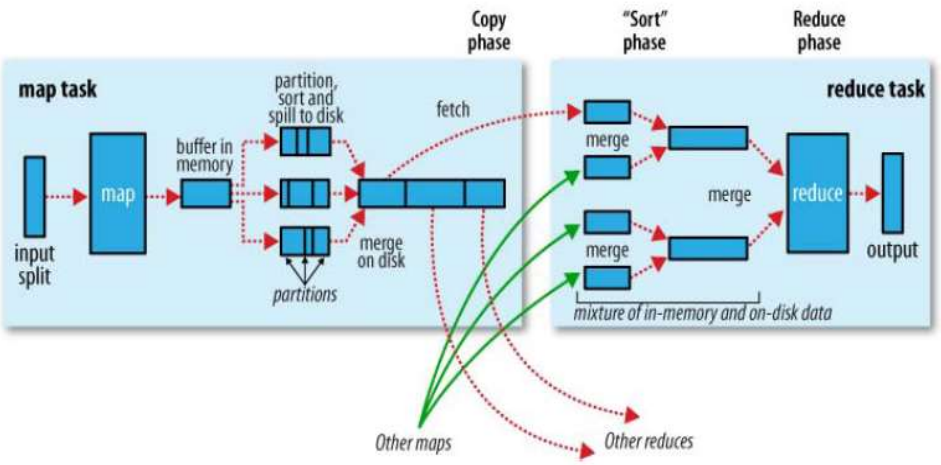
MapReduce Execution Details

- Single *master* controls job execution on multiple *slaves*
- Mappers preferentially placed on same node or same rack as their input block
  - Minimizes network usage
- Mappers save outputs to local disk before serving them to reducers
  - Allows recovery if a reducer crashes
  - Allows having more reducers than nodes

Shuffle & Sort

- The process between map and reduce
- Intermidiate output <k' , v'> are partitioned according to a ***partition function***
  - Usually a simple hash function for load balance
  - Specify your own if special purpose needed
- Exchange intermidiate output if needed
- Guarantees key order **within a reducer**

Shuffle & Sort



Shuffle & Sort: more details

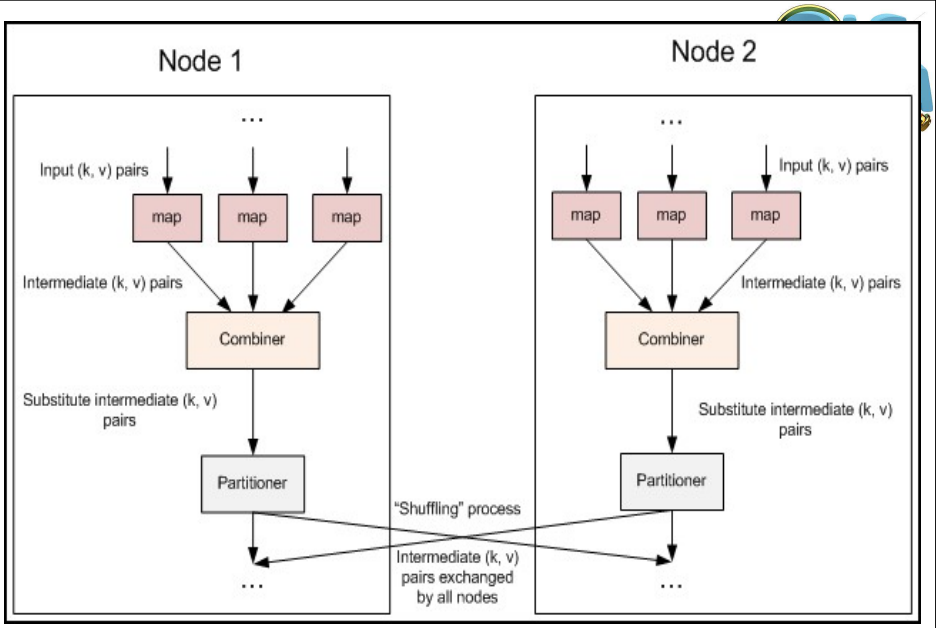
- Within a mapper:
  a. Keep emitting (k' , v') pairs to buffer until the **spill rate** of the buffer exceeds. After exceeding, the part of buffer is locked.
  b. An independent thread sorts the data within the locked buffer and **spill it out** to disk as a temporary document
  c. During the sorting, the mapper is only allowed to write to the remaining part of the buffer.
  d. After the map phase is done, combine the temporary documents into 1 document – the output of the mapper

Back to word count

- Consider "aaa aaa aaa aaa aaa bbb ccc…"
- Lots of (aaa, 1) are emitted by the mapper
- Extra overhead
  - Disk spill out
  - Network

The Combiner

- A pass executed between map and reduce
- A "mini-reduce" process that takes data from one machine only
  - But probably different from your reduce function
- To compress / trim the output from the map
- Optional: depends on your application
  - O: word count, min/max…
  - X: median

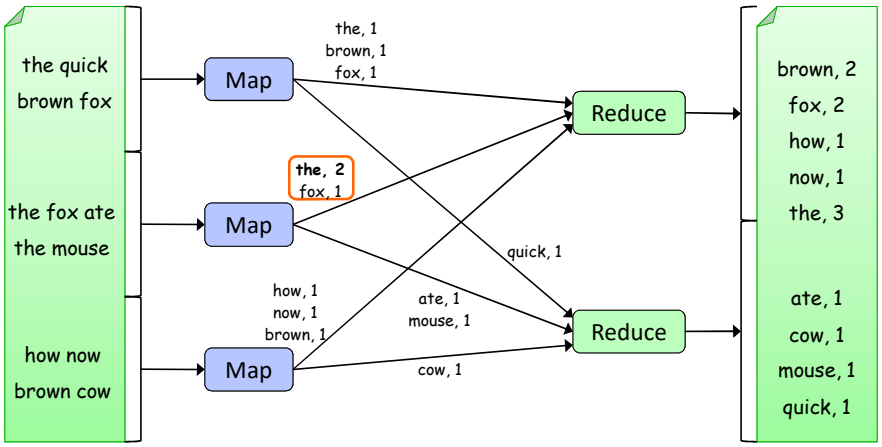


Combiner Example

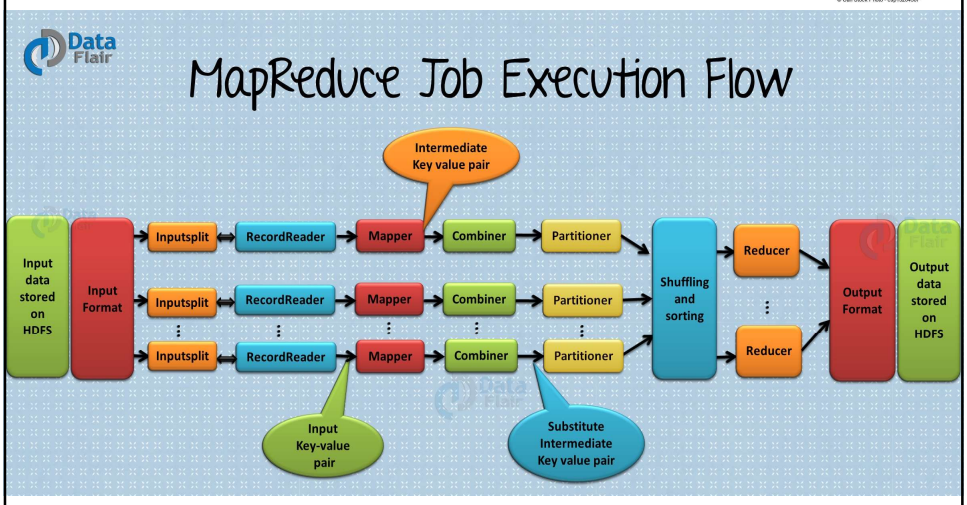
- A combiner is a local aggregation function for repeated keys produced by same map
- Works for associative functions like sum, count, max
- Decreases size of intermediate data
- Example: map-side aggregation for Word Count:

```
def combiner(key, values):
    output(key, sum(values))
```

Word Count with Combiner



MapReduce Job Execution Flow



MR Execution Summary 1

- One map task is created for each split which then executes map function for each record in the split.
- Multiple splits are processed in parallel.
- However, when splits are too small, the overhead of managing the splits and map task creation begins to dominate the total job execution time.
- For most jobs, a split size equals to the size of an HDFS block (64 MB, by default) is better.
- Execution of map tasks results into writing output to a local disk on the respective node and not to HDFS.
- Choosing local disk over HDFS is to avoid replication in case of HDFS store operation.

MR Execution Summary 2

- Map output is intermediate which is processed by reduce tasks to produce the final output.
- Once the job is complete, the map output can be thrown away. Storing it in HDFS with replication becomes overkill.
- On node failure, Hadoop reruns the map task on another node and re-creates the map output.
- An output of every map task is fed to the machine where reduce task is running.
- On this machine, the output is merged and then passed to the user-defined reduce function.
- Reduce output is stored in HDFS.

Advanced Issues: Scheduling

- One master, many workers
  - Input data split into M map tasks (typically 64 MB in size)
  - Reduce phase partitioned into R reduce tasks
  - Tasks are assigned to workers dynamically
  - Often: M=200,000; R=5,000; workers=2,000
- Master assigns each map task to a free worker
  - Considers locality of data to worker when assigning task
  - Worker reads task input (often from local disk!)
  - Worker produces R local files containing intermediate k/v pairs

Advanced Issues: Scheduling (cont.)

- Master assigns each reduce task to a free worker
  - Worker reads intermediate k/v pairs from map workers
  - Worker sorts & applies user's Reduce op to produce the output
- Fine granularity tasks: many more map tasks than machines
  - Minimizes time for **fault recovery**
  - Possible to have **pipelined shuffling** with map execution
  - Better **dynamic load balancing**
  - Why not as many map task as possbile?

Advanced Issues: Fault Tolerance

On worker failure:

- Detect failure via periodic **heartbeats**
- Re-execute completed and in-progress map tasks
- Re-execute in progress reduce tasks
- Task completion committed through master

On master failure:

- State is checkpointed to GFS: new master recovers & continues

Refinement: Backup Tasks

- **Problem**: Slow workers significantly lengthen completion time
  - Resouce contentions with other jobs
  - Bad disks and soft errors
  - Processor cache disabled
- Stragglers(流浪者) problem: a small number of mappers or reducers takes significantly longer than the others to complete
- **Solution**: Near end of phase, spawn backup copies of tasks

Refinement: Locality Optimazation

- Replicate input file blocks
- Split tasks into the size of a GFS block
- Map tasks scheduled to the same machine or same rack with the blocks of input data

  => Each job can be done on the same machine
- **Effect**: Thousands of machines read input at local disk speed
  - Without this, rack switches limit read rate

Refinement: Skipping Bad Records

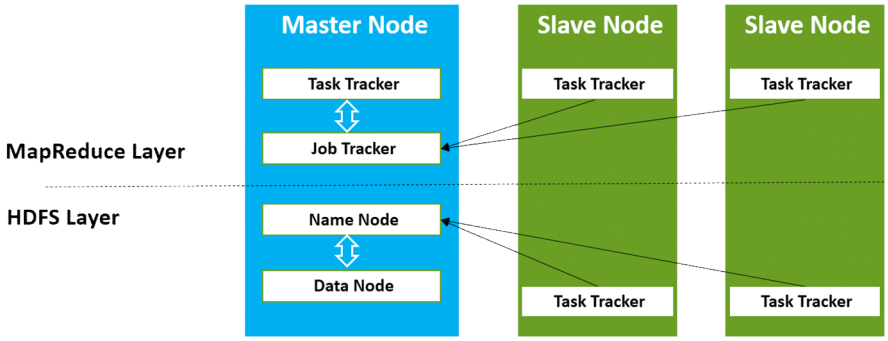
- **Problem**: Functions sometimes fail for particular inputs
- **Solution**: Skip them!
  - On seg fault, send UDP packet to inform master about which input caused the fault.
  - If master sees K failures for same record, skip the record afterwards

Implications for Multi-core Processors

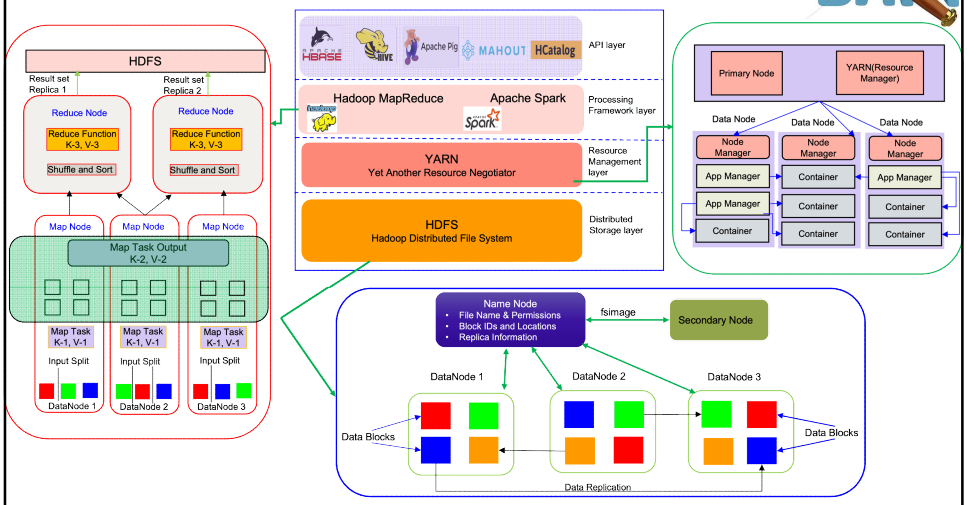
- Multi-core processors require parallelism
  - But many programmers are uncomfortable writing parallel programs
- MapReduce provides an easy-to-understand programming model for a very diverse set of computing problems
  - users don't need to be parallel programming experts
- Optimizations useful even in single machine, multi-core environment

Hadoop High-level Architecture

- Hadoop is based on two main components: MapReduce and HDFS.

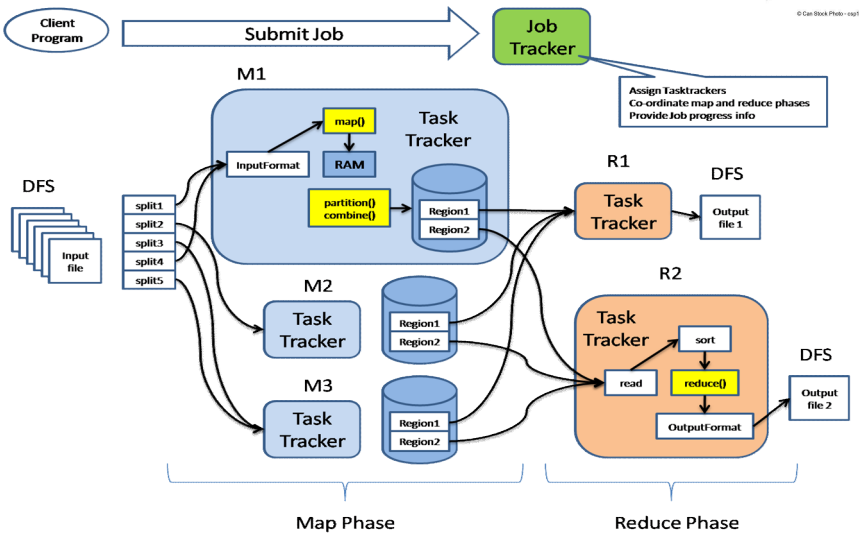


Hadoop in more details

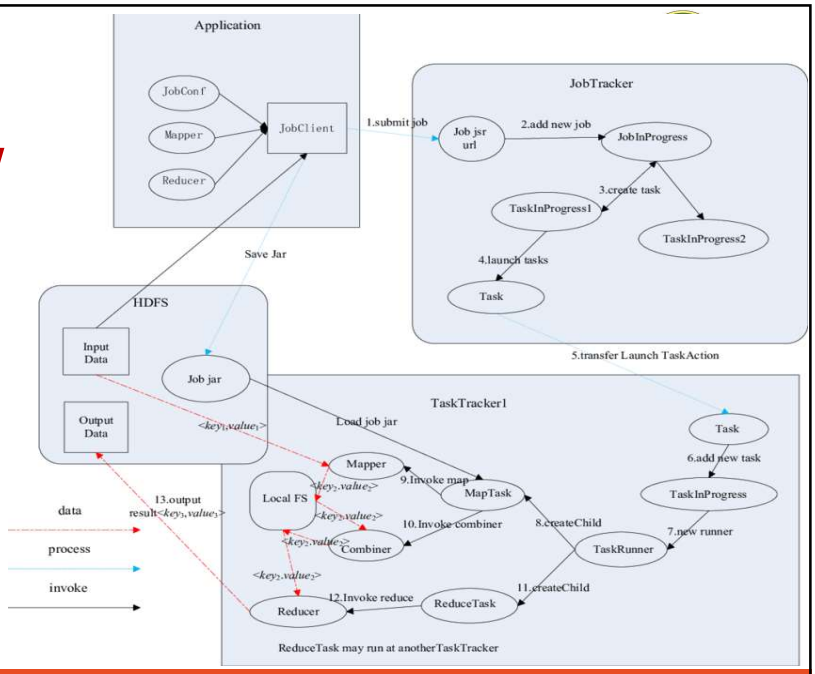


(More details in later lectures)

Hadoop/MR Architecture



MR Exec Flow

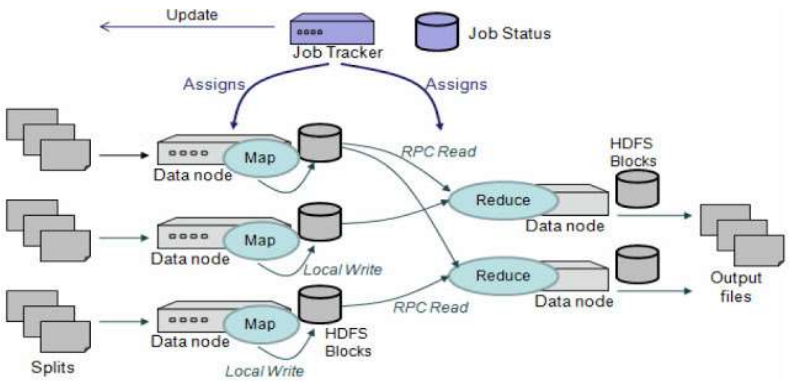


Problems with MapReduce

- It's hard & low-level for developers to write
  - Most developers are familiar with SQL
  - Solution: Apache Hive
- Expensive cost for fault recovery
  - Re-execute whole MR programs
  - Solution: Apache Spark's ***lineage***
- Requires intensive disk I/O
  - Intermidiate data is always written to local disk
  - Solution: Apache Spark's in-memory computing

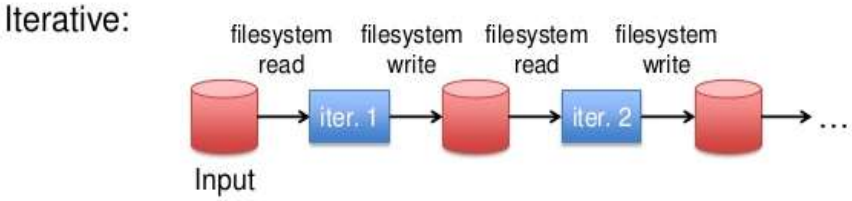
Problems with MapReduce

- MapReduce relies heavily on disk operations



Problems with MapReduce

- When doing iterative computation
  - Bad performance due to replication and disk I/O



In-Memory Computation is Faster

- Apache Spark in-memory computing
  - 10-100X faster than disk

Iterative:

Problems with MapReduce

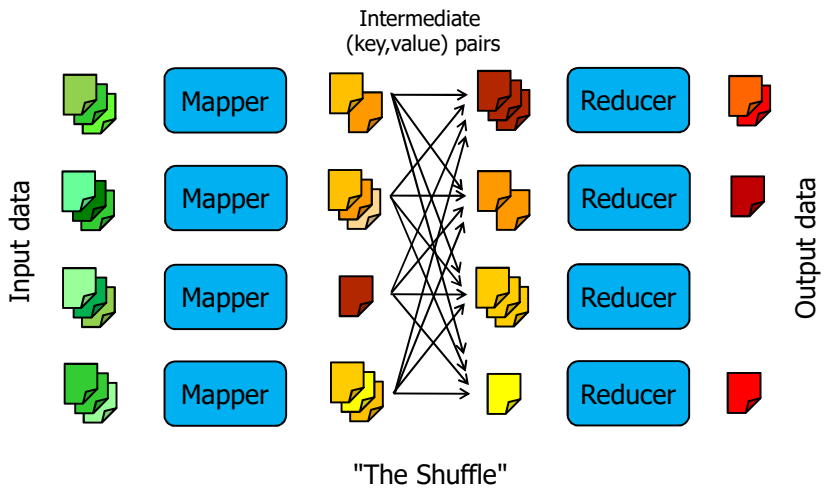
- Spawning each Mapper/Reducer takes time
  - Solution: Worker Pool (ex. Google Tenzing, an SQL query engine on Hadoop), it contains running processes as Mapper/Reducer
- Not very good for iterative graph computing
  - Solution: Google Pregel for large scale graph processing
- Not very good for interactive ad hoc queries
  - Solution: Google Dremel and BigQuery

Problem Solving with MapReduce

- How to design MapReduce algorithms for the following tasks
  - **Search**: Output lines matching certain patterns
  - **Sort**: Sorting numbers, words, …
  - **Inverted index**: build index from words to documents
  - **Data mining** algorithms (sequential pattern mining)
  - **BFS** on graph*
  - **PageRank***

MapReduce Algorithm Design

Recap: MapReduce Dataflow

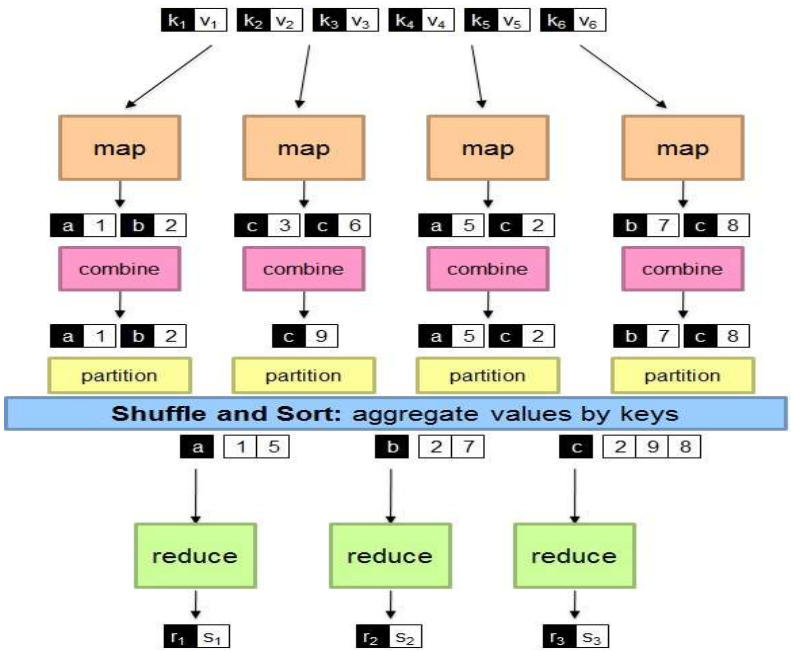


Recap: MapReduce

- Programmers must specify:
  map (k, v) → list(<k', v'>)
  reduce (k', list(v')) → <k'', v''>
  - All values with the same key are reduced together
- Optionally, also:
  partition (k', number of partitions) → partition for k'
  - Often a simple hash of the key, e.g., hash(k') mod n
  - Divides up key space for parallel reduce operations
  combine (k', v') → <k', v'>*
  - Mini-reducers that run in memory after the map phase
  - Used as an optimization to reduce network traffic
- The execution framework handles everything else…

"Everything Else"

- The execution framework handles everything else…
  - Scheduling: assigns workers to map and reduce tasks
  - "Data distribution": moves processes to data
  - Synchronization: gathers, sorts, and shuffles intermediate data
  - Errors and faults: detects worker failures and restarts
- Limited control over data and execution flow
  - All algorithms must expressed in m, r, c, p
- You don't know:
  - Where mappers and reducers run
  - When a mapper or reducer begins or finishes
  - Which input a particular mapper is processing
  - Which intermediate key a particular reducer is processing



Hadoop/MR on your Desk

- You can have a virtual Hadoop/MapReduce cluster easily with VM software such as VirtualBox.
- Download/install VirtualBox and configure a Linux VM(e.g. Ubuntu)
- Setting up a shared folder between host OS and VM is quite convenient for file transfer.
- On the VM, download/install Java, Hadoop and Python.
- There will be trouble ahead but the process is a good training!

Modes of Installation

- You may setup a Hadoop cluster in one of the three modes:
  - Local (Standalone) Mode
  - Pseudo-Distributed Mode
  - Fully-Distributed Mode
- If you are not familiar with Linux, start with the standalone mode.
- If you were a Linux guru, you may setup three or more VMs and take on the fully-distributed mode directly.

Recap: Word Count

These types depend on the input data

```
map(key:URL, value:Document)
{
  String[] words = value.split(" ");
  foreach w in words
    emit(w, 1);
}

reduce(rkey:String, rvalues:Integer[])
{
  Integer result = 0;
  foreach v in rvalues
    result = result + v;
  emit(rkey, result);
}
```

Produces intermediate key-value pairs that are sent to the reducer

These types can be (and often are) different from the ones in map()

reduce gets all the intermediate values with the same rkey

Both map() and reduce() are stateless: Can't have a global variable that is preserved across invocations!

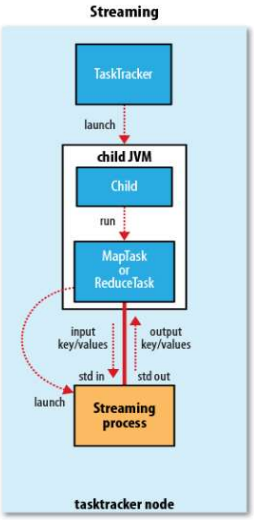
Any key-value pairs emitted by the reducer are added to the final output

MapReduce in Python

- Hadoop is written in Java. It is only nature that many MR programs are written in Java.
- Hadoop MR programs can be written in languages such as Python, C++, Ruby, etc.
- Traditionally, a Python code is translated using Jython into a Java jar file for execution.
- However, this is not very convenient and surely not very **Pythonic**!
- We will show you another way of writing Python MR code with Hadoop Streaming.

Hadoop Streaming API

- With streaming API, you can write MR program in all languages that can read/write standard I/O.
- Both the mapper and reducer use stdin/stdout for reading/writing.
- Mappers read from stdin for input data and print results to stdout.
- Reducers read mapper output from stdin and print results to stdout.



mapper.py

```
#!/usr/bin/env python3
# import sys to read/write data from/to STDIN/STDOUT
import sys
for line in sys.stdin:    # input from STDIN
    line = line.strip()   # rm leading/trailing whitespace
    words = line.split() # split the line into words
    for word in words:    # each with a count of 1
        # write the results to STDOUT;
        # will be the input of the reducer.py
        # tab-delimited; word appears once(1)
        print("%s\t%s" % (word, 1))
```

Data File and Execution

```
host:~/Documents$ cat wc_data.txt
NDHU CSIE is the best CS department
Welcome to CSIE for CS degrees
host:~/Documents$ cat wc_data.txt | python3 mapper.py
```

```
NDHU     1
CSIE     1
is       1
the      1
best     1
CS       1
department      1
Welcome  1
to       1
CSIE     1
for      1
CS       1
degrees  1
```


reducer.py(1)

```
#!/usr/bin/env python3
import sys
# dictionary to map words to counts
wordcount = {}

# input comes from STDIN
for line in sys.stdin:
    # rm leading and trailing whitespace
    line = line.strip()
```

reducer.py(2)

```
# slit the input from mapper.py into 2 elements
word, count = line.split('\t', 1)
# convert count (currently a string) to int
try:
    count = int(count)
except ValueError:  # simply ignore if err
    continue

try:  # accumulate the count
    wordcount[word] = wordcount[word] + count
except:
    wordcount[word] = count
```


reducer.py(3)

```
# write the tuples to stdout
for word in wordcount.keys():
  print("%s\t%s" % (word, wordcount[word]))
```

How to Submit ?

run.sh

```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -D mapred.reduce.tasks=2 -input /user/showyang/wc_data.data  -output /user/showyang/output -mapper "python3 mapper.py" -reducer "python3 reducer.py" -file mapper.py -file reducer.py
```

- Programs are local files.
- Input/output files are on the HDFS.
- Paths may be different based on your system.
- Make it work and send me video to demonstrate.

MapReduce Commands

- You can also invoke all mapreduce commands by the bin/mapred script.

```
mapred [SHELL_OPTIONS] COMMAND
  [GENERIC_OPTIONS] [COMMAND_OPTIONS]
```

- The streaming command looks like:

```
mapred streaming [genericOptions]
      [streamingOptions]
```

- Try submitting your Python Word Count job with the mapred script. (exercise)

MapReduce Jobs

- Tend to be very short, code-wise
  - IdentityReducer is very common
- "Utility" jobs can be composed
- Represent a *data flow*, more so than a procedure

MR Algorithm Design Principles

- Think data, not flow!
- Decompose the problem into modules connected by data flow, not algorithmic flow.
- Design MR job(s) for each module.
- Link jobs with (key, value) pairs.
- Key and value can potentially be anything !!

Sorting

Inputs:

- A file of values to sort, one value per line.
- Mapper key is file ID, line number
- Mapper value is the contents of the line

- This can be easily generalized into sorting multiple files. (Exercise)

Sort Algorithm

- Takes advantage of reducer properties: (key, value) pairs are processed in order by key; reducers are themselves ordered

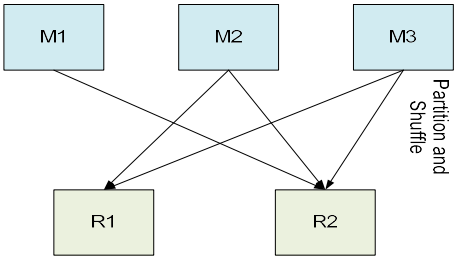
- Mapper: Identity function for value

  (k, v) → (v, _)

- Reducer: Identity function (k', _) -> (k', “”)

Sort: The Trick

- (key, value) pairs from mappers are sent to a particular reducer based on hash(key)
- Must pick the hash function for your data such that $k_1 < k_2 \Rightarrow hash(k_1) < hash(k_2)$



Searching

- Given a set of files containing lines of text and a search pattern to find.
- Determine the files that matches the pattern.
- Can be easily generalized into determining (file, [l1, l2, …]), i.e. all lines that match the pattern. (Exercise)
- Search pattern sent as special parameter

Search Algorithm

- Mapper:
  - Given (fileID, some text) and "pattern", if "text" matches "pattern" output (filename, _)
- Reducer:
  - Identity function

Search: An Optimization

- Once a file is found to be interesting, we only need to mark it that way once
- Use *Combiner* function to fold redundant (filename, _) pairs into a single one
  - Reduces network I/O

Inverted Index

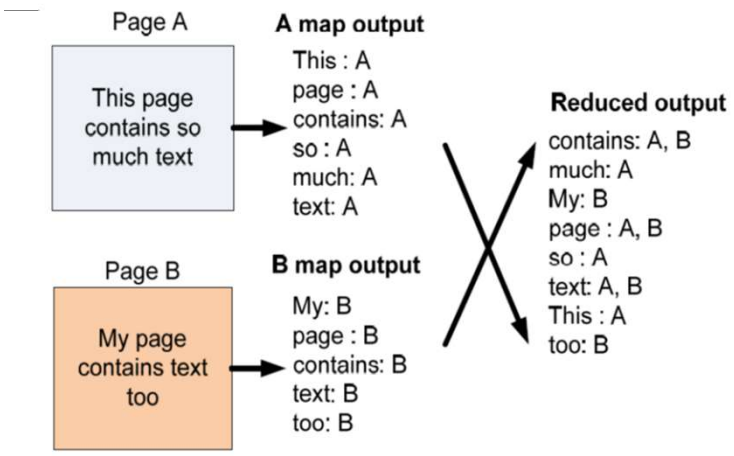
- Given a set of document(text) files.
- For each word, determine the docs in which the word appears. (Boolean)
- For each word, determine the docs & positions where the word appears. (Exercise)

Inverted Index Algorithm

- Mapper key is file ID
- Mapper value is the contents of the file.

- Mapper: For each word in (fileID, words), map to (word, fileID)
- Reducer: For each word, output (word, [f1, f2, …])

Indexing: Data Flow



TF-IDF

- Term Frequency – Inverse Document Frequency
  - Relevant to text processing
  - Common web analysis algorithm
  - To determine the importance of a term within a corpus (set of docs).

The Algorithm, Formally

$$\mathrm{tf_i} = \frac{n_i}{\sum_k n_k}$$

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$\mathrm{tfidf} = \mathrm{tf} \cdot \mathrm{idf}$$

- $|D|$: total number of documents in the corpus
- $|\{d : t_i \in d\}|$: number of documents where the term t_i appears (that is $n_i \neq 0$).

Information We Need

- Number of times term X appears in a given document (n_i)
- Total number of terms in each document ($\Sigma_k n_k$)
- Number of documents X appears in ($|\{d : t_i \in d\}|$)
- Total number of documents ($|D|$)

Job 1: Word Frequency in Doc

- Mapper
  - Input: (docname, contents)
  - Output: ((word, docname), 1)
- Reducer
  - Sums counts for each word in document
  - Outputs ((word, docname), *n*)
- Combiner is the same as Reducer

Job 2: Total Word Counts For Docs

- Mapper
  - Input: ((word, docname), *n*)
  - Output: (docname, (word, *n*))
- Reducer
  - Sums frequency of individual *n*'s in same doc
  - Feeds original data through
  - Outputs ((word, docname), (*n*, *N*))

Job 3: Word Frequency In Corpus

- Mapper
  - Input: ((word, docname), (*n, N)*)
  - Output: (word, (docname, *n*, *N*, 1))
- Reducer
  - Sums counts for word in corpus
  - Outputs ((word, docname), (*n*, *N*, *m*))

Job 4: Calculate TF-IDF

- Mapper
  - Input: ((word, docname), (*n*, *N*, *m*))
  - Assume D is known (or, easy MR to find it, exercise!)
  - Output ((word, docname), TF*IDF)
- Reducer
  - Just the identity function

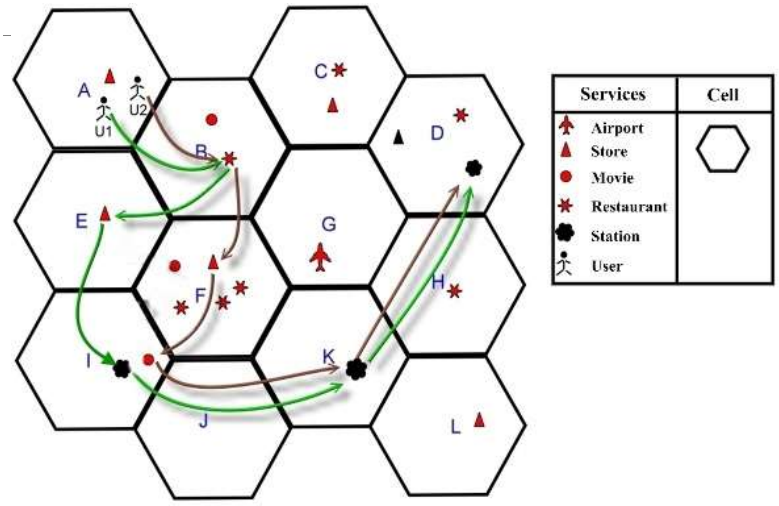
Working At Scale

- Buffering (doc, *n*, *N*) counts while summing 1's into *m* may not fit in memory
  - How many documents does the word "the" occur in?
- Possible solutions
  - Ignore very-high-frequency words (AKA stop words)
  - Write out intermediate data to a file
  - Use another MR pass

Final Thoughts on TF-IDF

- Several small jobs add up to full algorithm
- Lots of code reuse possible
  - Stock classes exist for aggregation, identity
- Jobs 3 and 4 can really be done at once in same reducer, saving a write/read cycle
- Very easy to handle medium-large scale, but must take care to ensure flat memory usage for largest scale

Sequential Activity Mining in Mobile Environments



Sequential Activity Mining

Data Base D

TID	Behavior Items
100	A1,B2,C3,E7
200	A1,B2,E3,E5
300	A1,B2,G3
400	A1,E8

A1

Items	Count
B2	3
A1	4
E	4
3	3

C2

Items	Count	Items	Count
(B2,A1)	0	(E,A1)	0
(B2,E)	2	(E,3)	0
(B2,3)	3	(3,B2)	0
(A1,B2)	3	(3,A1)	0
(A1,E)	3	(3,E)	2
(A1,3)	3	(E,B2)	0

A2

Items	Count
(B2,E)	2
(B2,3)	3
(A1,B2)	3
(A1,E)	3
(A1,3)	3
(3,E)	2

C3

Items	Count
(B2,3,E)	2
(A1,B2,E)	2
(A1,B2,3)	3
(A1,3,E)	2

A3

Items	Count
(B2,3,E)	2
(A1,B2,E)	2
(A1,B2,3)	3
(A1,3,E)	2

C4

Items	Count
(A1,B2,3,E)	2

A4

Items	Count
(A1,B2,3,E)	2

support=2

Ai : Length i large Activity Set.
Ci : Length i condidate Set.

Activity Mining V1

- Direct conversion from the original algorithm.
- **Job1**: Large-1 activity set generation.
  - Mapper: given (TID, Behavior_Items), generate all possible (item, 1).
  - Reducer: for each item, sum the count and emit all ((item), n) for items with n >= support.

Activity Mining V1

- **Job2**: Large-2 ~ n activity set generation.
- Mapper: Given a (activityset, count) pair (A, n), emit (prefix of A except the last item, last item).
- Reducer: Given (prefix, [m1, m2, …]), emit all possible ((prefix, mi, mj), _) as candidates.
- **Job3**: Given each candidate pattern, count frequency in Transaction Data Base (D) and keep only those with enough support. (Exercise)

Activity Mining V1

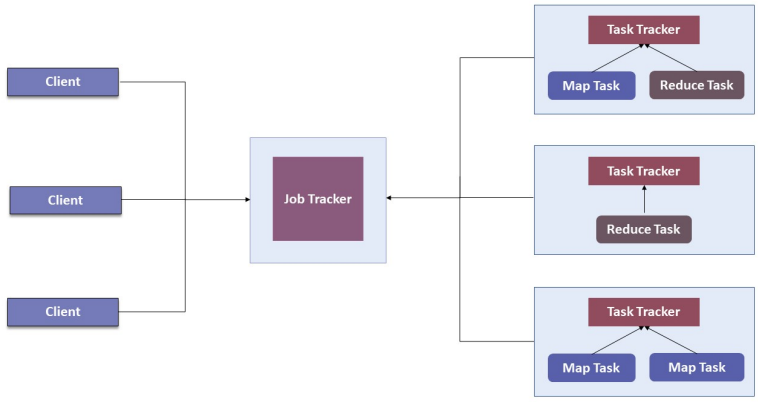
- The final result is the union of the output of all reducers that generate (ActivitySet, count)
- Optimization: (Exercise)
  - Use combiner to compute local sums.
  - Use more than one reducer for candidate generation.
  - Use efficient MR for Transaction DB scanning.

Activity Mining V2

- One pass MapReduce algorithm
- Mapper: Given (TID, Behavior_Items), generate all possible (pattern, 1).
- Reducer: Sum the count for the same pattern and emit (pattern, n) if n >= support.
- That's it !!
- Optimization: (Exercise)
  - Parallelize the mapper?
  - Stop counting whenever n>= support?

Hadoop 1 (MRV1) Revisit

- Hadoop version 1.0 is referred as MARV1(MapReduce v1)



Problems with MRV1

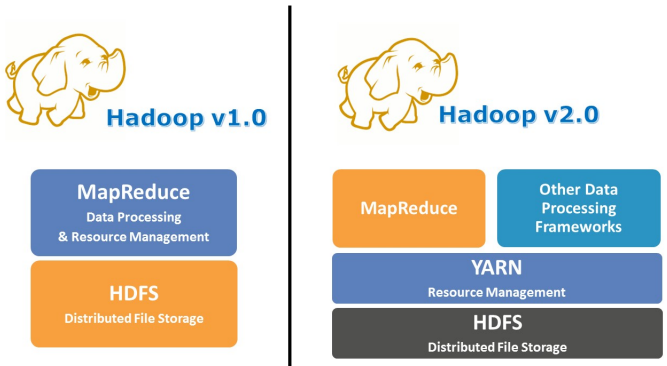
- Only for batch processing, not real-time processing
- MRV1 & HDFS support up to 4000 nodes/cluster
- JobTracker's load too heavy, single point of failure
- NameNode's load too heavy, single point of failure
- Scalability issues due to problems above
- No Multi-tenancy support
- Only run MapReduce jobs, can not support other frameworks
- Utilization of resources is inefficient

Hadoop v2.0

- Introducing YARN(Yet Another Resource Negotiator), a resource management system for Hadoop (also known as MapReduce 2 or MRV2)
- Act as a connecting link between high level applications and low level Hadoop environment
- Transform Hadoop from only MapReduce framework to big data processing core
- Scale much better than Hadoop 1.

Hadoop 1 vs Hadoop 2

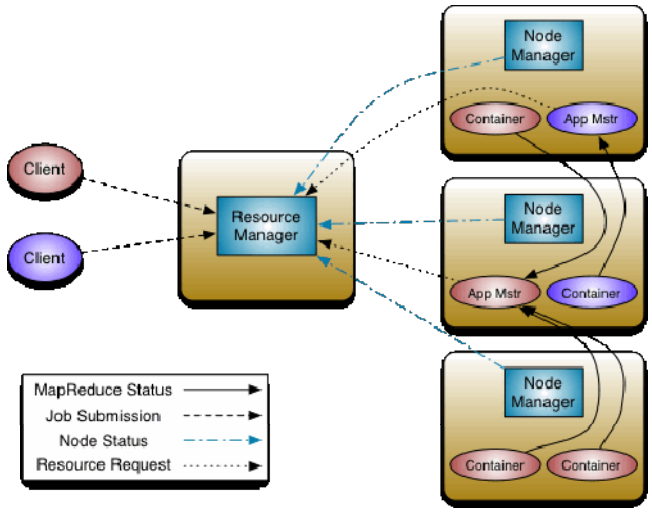
- Hadoop 2 offers better performance, scalability, fault-tolerance and multiple processing frameworks.



YARN Components

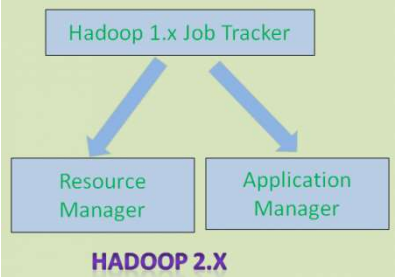
- **Resource Manager**: Runs on a master daemon and manages resources across the cluster.
- **Node Manager**: Run on the slave daemons and are responsible for the execution of a task on every single Data Node.
- **Application Master**: Manages the user job lifecycle and resource needs of individual applications. It works along with the Node Manager and monitors the execution of tasks.
- **Container**: Collection of resources such as RAM, CPU cores, Network, HDD etc on a single node.

Hadoop YARN Architecture

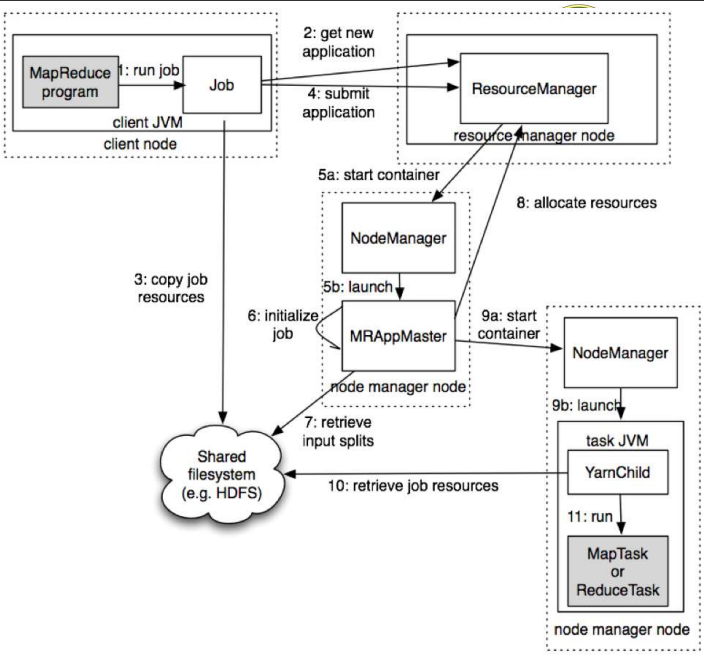


YARN Benefits

- Split Job Tracker into separate Resource Manager and Application Manager (more later)
- Benefits:
  - Highly scalability
  - Highly availability
  - Supports multiple programming models
  - Supports multi-tenancy
  - Supports multiple namespaces
  - Improved cluster utilization
  - Improve horizontal scalability



How does YARN (MRV2) works?



Failure Recovery in YARN

- Task failure, same as in MapReduce 1
- Application Master failure:
  - Resource Manager notices failed AppMaster
  - Resource Manager starts a new instance of AppMaster in new container
  - Client experiences a timeout and get a new address of AppMaster from ResourceManager
- Resource Manager failure:
  - Resource Managers have checkpointing mechanism which saves its state to persistent storage.
  - After crash, administrator brings new Resource Manager up and it recovers saved state.

What's New in Hadoop 3?

3.0 Key Advancements

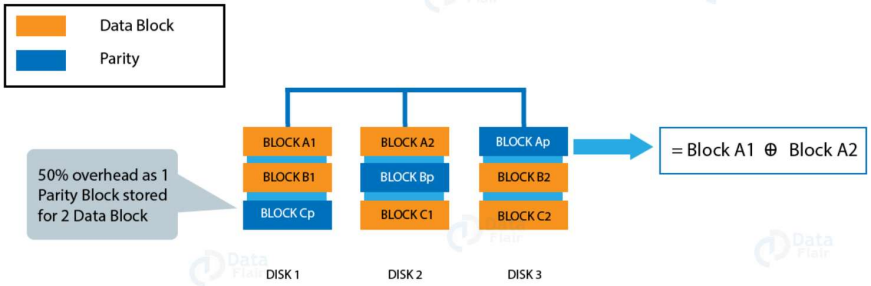
- Erasure Coding: Introduce the concept of erasure coding, a more **storage-efficient** alternative to traditional replication.
- Improved Resource Utilization: Introduce containerization with Docker and Kubernetes, allowing for better isolation, resource allocation, and efficient utilization of cluster resources.
- Enhanced Data Processing Engines: Introduce improvements to existing data processing engines like MapReduce and Hive. It also provide better integration and support for newer engines like Apache Spark and Apache Flink, enabling faster and more flexible data processing workflows.

3.0 Key Advancements

- Namenode Federation and High Availability: Address the scalability and availability challenges of the Namenode by introducing Namenode Federation and the standby Namenode. This improved fault tolerance, reduced downtime, and enhanced the overall reliability of the Hadoop cluster.
- GPU Support: Embrace the power of GPUs by introducing support for GPU acceleration. This enabled organizations to leverage GPU resources for parallel processing and achieve significant performance gains in data-intensive tasks like machine learning and deep learning.

Erasure Coding

- Hadoop 2.x uses replication (default 3). Storage overhead is 200%.
- Erasure coding stores 1 parity block for 2 data blocks. Same level of fault tolerance with 50% overhead.



Some Other 3.0 Features

- Opportunistic Containers have low priority than Guaranteed containers and wait at the NodeManager when no resources is available.
- Distributed Scheduling incorporates opportunistic containers for more flexible scheduling.
- Support for more than two NameNodes made the system more highly available.
- Intra-DataNode Balancer balances the disk load in a DataNode.(**HDFS balancer** addresses only **internode** data skew)

Hadoop 3.x vs 2.x

- Should use Hadoop 3 whenever possible!!

Features	Hadoop 2.x	Hadoop 3.x
Min Java Version Required	Java 7	Java 8
Fault Tolerance	Via replication	Via erasure coding
Storage Scheme	3x replication factor for data reliability, 200% overhead	Erasure coding for data reliability, 50% overhead
Yarn Timeline Service	Scalability issues	Highly scalable and reliable
Standby NN	Supports only 1 SBNN	Supports only 2 or more SBNN
Heap Management	We need to configure HADOOP_HEAPSIZE	Provides auto-tuning of heap
File System Compatability	HDFS, FTP, S3, Windows Azure Storage Blobs	Support all file systems

Some HDFS Commands

- Create a directory in HDFS

hdfs dfs -mkdir /home/hadoop/dir1

- List the content of a directory

hdfs dfs -ls /home/hadoop

- Upload and download a file in HDFS

hdfs dfs -put file.txt /home/hadoop/dir1/

hdfs dfs -get /home/hadoop/dir1/file.txt /home/hadoop

- Look at the content of a file

hdfs dfs -cat /home/hadoop/dir1/book.txt

- Many more commands, similar to Linux

Assignment 1

- Test run the word count example. (No need to turn in anything.)
- Implement the Sorting algorithm.
- Implement the Searching algorithm.
- Implement the TF-IDF computation algorithm.
- Implement the Activity Mining algorithms. (optional)
- Due date: three weeks!