CS143: Map Reduce (Spark)

Book Chapters

(7th) Chapters 10.3-4

Distributed Computing on Cluster

- Often, our data is non-relational (e.g., flat file) and huge
  - Billions of query logs
  - Billions of web pages
  - ...
- Q: Can we perform analytics on large data quickly using thousands of machines? How can we help programmer write parallel code running in distributed clusters?

Examples

Example 1: Search log analysis

- Log of billions of queries. Count frequency of each query

```
Input query log:
    cat,time,userid1,ip1,referrer1
    dog,time,userid2,ip2,referrer2
    ...
Output query frequency:
    cat 200000
    dog 120000
    ...
```

  - Log file is spread over many machines
- Questions
  - Q: How can we do this?
  - Q: How can we run it on thousands of machines in parallel?
    - Q: Can we process each query log entry independently?
    - Q: How can we combine the results?

Example 2: Web Indexing

- 1 billion pages. build inverted index

```
Input documents:
    1: cat chases dog
    2: dog hates zebra
    ...
Output index:
    cat 1,2,5,10,20
    dog 2,3,8,9
    ...
```

- Questions
  - Q: How can we do this?
  - Q: How can we run it on thousands of machines?
    - Q: Can we process each page independently?
    - Q: How can we aggregate extracted (word, docid)'s?

Generalization of Examples

- Common pattern in the two examples
  - Input data consists of multiple independent units
    - Each line of query log
    - Each web page
  - Partition input data into multiple "chunks" and distribute them to multiple machines
  - Transformation/map input into (key, value) tuples
    - Query log: query_log_line → (query, 1)
    - Indexing: web_page → (word1, page_id), (word2, page_id), ...
  - Reshuffle tuples of the same key to the same machine
  - Aggregate/reduce the tuples of same keys
    - Query log: (query, 1), (query, 1), ... → (query, count)
    - Indexing: (word, 1), (word, 3), ... → (word, [1, 3, ...])
  - Collect and output the aggregation results
- The two examples are almost the same except
  - "The mapping function"
    - Query log: query_log_line → (query, 1)
    - Indexing: web_page → (word1, page_id), (word2, page_id), ...
  - "The reduction function"
    - Query log: (query, 1), (query, 1), ... → (query, count)
    - Indexing: (word, 1), (word, 3), ... → (word, [1, 3, ...])

Map/Reduce Programming Model

- Many data processing jobs can be done as a sequence of
  1. Map: $(k, v) \rightarrow (k', v'), (k'', v''), ...$
  2. Reduce: partition/group by k and "aggregate" v's of the same k

  - Output of map function depends only on the input (k, v), not any other input
    * Each map task can be executed independently of others
  - Reduction on different keys are independent of each other
    * Each reduce task can be executed independently of others
    * Reduce function should be agnostic to the order of v's
- If any data processing follows the previous pattern, they can be parallelized by
  1. Split the input into independent chunks
  2. Run "map" tasks on the chunks in parallel on multiple machines
  3. Partition the output of the map task by the output key
  4. Move data of the same partition to the same node
  5. Run one reduce task per each partition

  - Only the map and reduce functions are different per app
- Under Map/Reduce programming model:
  - Programmer provides
    * Map function $(k, v) \rightarrow (k', v')$
    * Reduce function $(k, [v1, v2, ...]) \rightarrow (k, aggr([v1, v2, ...]))$
  - MapReduce handles the rest
    * Automatic data and task, partition, distribution, and collection
    * Failure and speed disparity handling

Systems

Hadoop

- First open source implementation of GFS (Google File System) and MapReduce
  - Implemented in Java
- Map and reduce functions are implemented by:
  - `Mapper.map(key, value, output, reporter)`
  - `Reducer.reduce(key, value, output, reporter)`

Spark

- Open source cluster computing infrastructure
- Supports MapReduce and SQL
  - Supports data flow more general than simple MapReduce
- Input data is converted into RDD (resilient distributed dataset)

  - A collection of independent tuples
  - The tuples are automatically distributed and shuffled by Spark

- Supports multiple programming languages

  - Scala, Java, Python, ...
  - Scala and Java are much more performant than others

Spark: Example

- Count words in a document

```
lines = sc.textFile("input.txt")
words = lines.flatMap(lambda line: line.split(" "))
word1s = words.map(lambda word: (word, 1))
wordCounts = word1s.reduceByKey(lambda a,b: a+b)
wordCounts.saveAsTextFile("output")
```

  - `map()`: one output per one input
  - `flatMap()`: multiple outputs per one input

Key Spark Functions

- Transformation: Convert RDD tuple into RDD tuple(s)

  - `map()`: convert one input tuple into one output tuple
  - `flatMap()`: convert one input into multiple output tuples
  - `reduceByKey()`: specify how two input "values" should be aggregated
  - `filter()`: filter out tuples based on condition

- Action: Perform "actions" on RDD

  - `saveAsTextFile()`: save RDD in a directory as text file(s)
  - `collect()`: create Python tuples from Spark RDD
  - `textFile()`: create RDD from text (each line becomes an RDD tuple)