Iterative Algorithms on hadoop

15-440 Fall 22
Ammar Karkour

How can different Hadoop jobs communicate?

Through hdfs

1. Job at Iteration i:
   a. Reads it's output from all different output files ***part-0000z***
   b. Writes **what it wants to send to Job at iteration i+1** to file ***x***

2. Job at iteration i+1:
   a. Mapper: reads data from file ***x***

How to read output from all different output files?

```
PathFilter filter = new PathFilter () {
        public boolean accept(Path file) {
                return file.getName().startsWith("part-");
        }
};


FileStatus[] output_files = hdfs.listStatus(output_file, filter);

reader = new BufferedReader(new InputStreamReader(hdfs.open(output_files[i].getPath())));

String read_line = reader.readLine();

……
```

How does a mapper read from file ***x***?

There are different ways of doing it:

- Option 1: define function **public void configure(JobConf job) { … } or protected void setup(org.apache.hadoop.mapreduce.Mapper.Context context)** in the Mapper class
  - These function run once at the beginning of each task. So you would read file ***x*** there and assign the value in it to a global variable that can be accessed by the **map** function.

- Option 2: Just read the value in the m function map itself

Useful links

- Hadoop docs 1
- Hadoop docs 2