Hadoop BLAST Project

Abderrezak Allalen

Varun Patil

M.S. Data Science

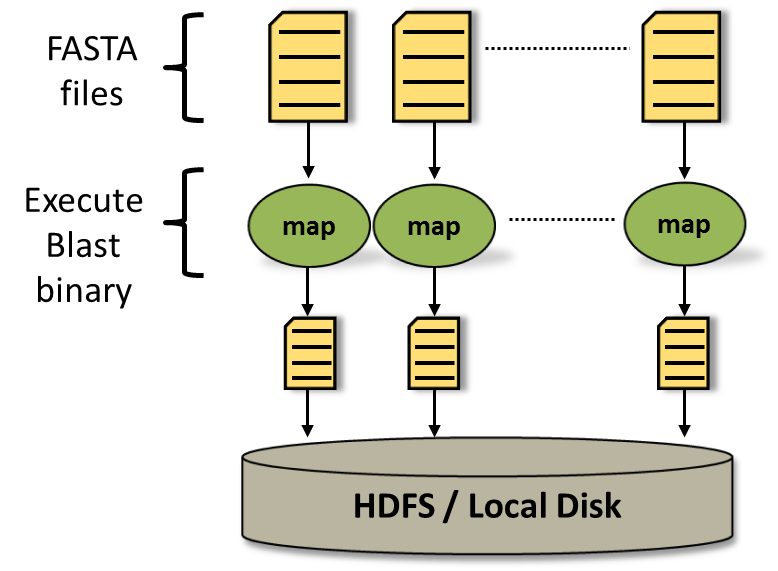
Indiana University

02/28/2015

BLAST short for Basic Local Alignment Search Tool is a widely used algorithm for sequence alignment. BLAST is one of the commonly used techniques to solve scientific problems in the bioinformatics discipline.

In this project a parallel version of BLAST is implemented using Hadoop MapReduce framework. The application is written in Hadoop Map only in form of RunnerMap.java.

**Hadoop Blast Data Flow:**



**Description of Data flow:**

The DataAnalysis.java class is the main class which sets Hadoop job configuration.

In the RunnerMap.java class, the map method sets up the distributed cache and generates two absolute locations for BLAST binary and BLAST database

*this.localDB = local[0].toUri().getPath() + "/" + conf.get(DataAnalysis.DB\_ARCHIVE) + "/" + conf.get(DataAnalysis.DB\_NAME);  
this.localBlastProgram = local[0].toUri().getPath();*

It then copies the assigned FASTA files to local disk.

*fs.copyToLocalFile(inputFilePath, new Path(localInputFile));*

The BLAST binary is executed with the assigned FASTA files which are then passed as key value pair <filename, filepath on HDFS>.

*String execCommand = cmdArgs.replaceAll("#\_INPUTFILE\_#", localInputFile);*

*if (cmdArgs.indexOf("#\_OUTPUTFILE\_#") > -1) {*

*execCommand = execCommand.replaceAll("#\_OUTPUTFILE\_#", outFile);*

*}else{*

*outFile = stdOutFile;*

*}*

*endTime = Double.toString(((System.currentTimeMillis() - startTime) / 1000.0));*

*System.out.println("Before running the executable Finished in " + endTime + " seconds");*

*execCommand = this.localBlastProgram + File.separator + execName + " " + execCommand + " -db " + this.localDB;*

The output FASTA files of BLAST binary will be uploaded back to HDFS.

Path outputDirPath = new Path(outputDir);

Path outputFileName = new Path(outputDirPath,fileNameOnly);

fs.copyFromLocalFile(new Path(outFile),outputFileName);

There is no need of Reduce function since the Map function just compares the FASTA files with the BLAST database and returns the similar protein sequences.