Blast – sequence alignment

Create a bunch of map tasks and execute the input files.

No reduction

Distributed cache mechanism - important

Report – 5 questions.

Custom input format – find the path, hand it to external program, get the output.

Runner map – pick things from hdfs and run external program

Outputhandler – gets output from eternal program and write it to the disk

Data file input format – custom

Demo:

.sh –

Data analysis . java –

Arguments:

0 – zip file

1 – external executable

2 – working directory

4 – database name

5 – input dir

6 – output dir.

Runner map:

Where to find external executable.

Where to find input files.

* What is Hadoop Distributed Cache and how is it used in this program?
* Write the two lines that put and get values from Distributed cache. Also include the method and class information.
* In previous projects we used Hadoop’s TextInputFormat to feed in the file splits line by line to map tasks. In this program, however, we want to feed in a whole file to a single map task. What is the technique used to achieve this? Also, briefly explain what are the key and value pairs you receive as input to a map task and what methods are responsible for producing these pairs?
* Do you think this particular implementation will work if the input files are larger than the default HDFS block size? Briefly explain why. [Hint: you can test what will happen by concatenating the same input file multiple times to create a larger input file in the resources/blast\_input folder]
* If you wanted to extend this program such that all output files will be concatenated into a single file, what key and value pairs would you need to emit from the map task? Also, how would you use these in the reduce that you would need to add?

Report:

1. Google.. (storage)
2. Code
3. Read path to given file.