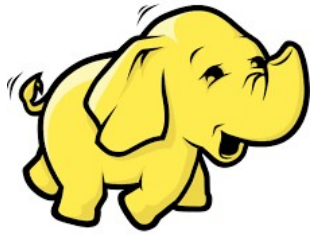
Hadoop Streaming

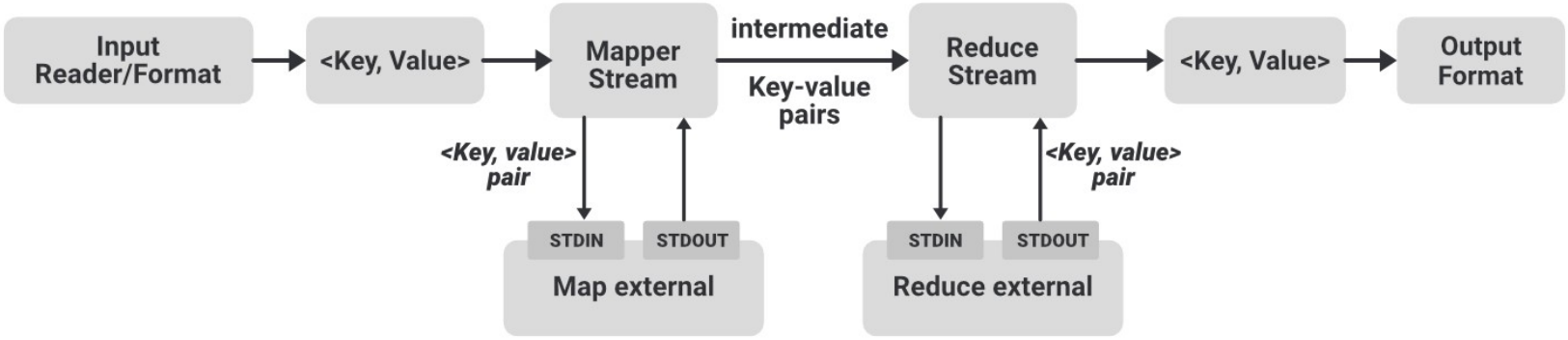
Tushar B. Kute,
http://tusharkute.com



Hadoop Streaming

- It is a utility or feature that comes with a Hadoop distribution that allows developers or programmers to write the Map-Reduce program using different programming languages like Ruby, Perl, Python, C++, etc.
- We can use any language that can read from the standard input(STDIN) like keyboard input and all and write using standard output(STDOUT).
- We all know the Hadoop Framework is completely written in java but programs for Hadoop are not necessarily need to code in Java programming language. feature of Hadoop Streaming is available since Hadoop version 0.14.1.

Hadoop Streaming



Hadoop Streaming

- In the above example image, we can see that the flow shown in a dotted block is a basic MapReduce job. In that, we have an Input Reader which is responsible for reading the input data and produces the list of key-value pairs.
- We can read data in .csv format, in delimiter format, from a database table, image data(.jpg, .png), audio data etc.
- The only requirement to read all these types of data is that we have to create a particular input format for that data with these input readers.

tusharkute.com

Hadoop Streaming

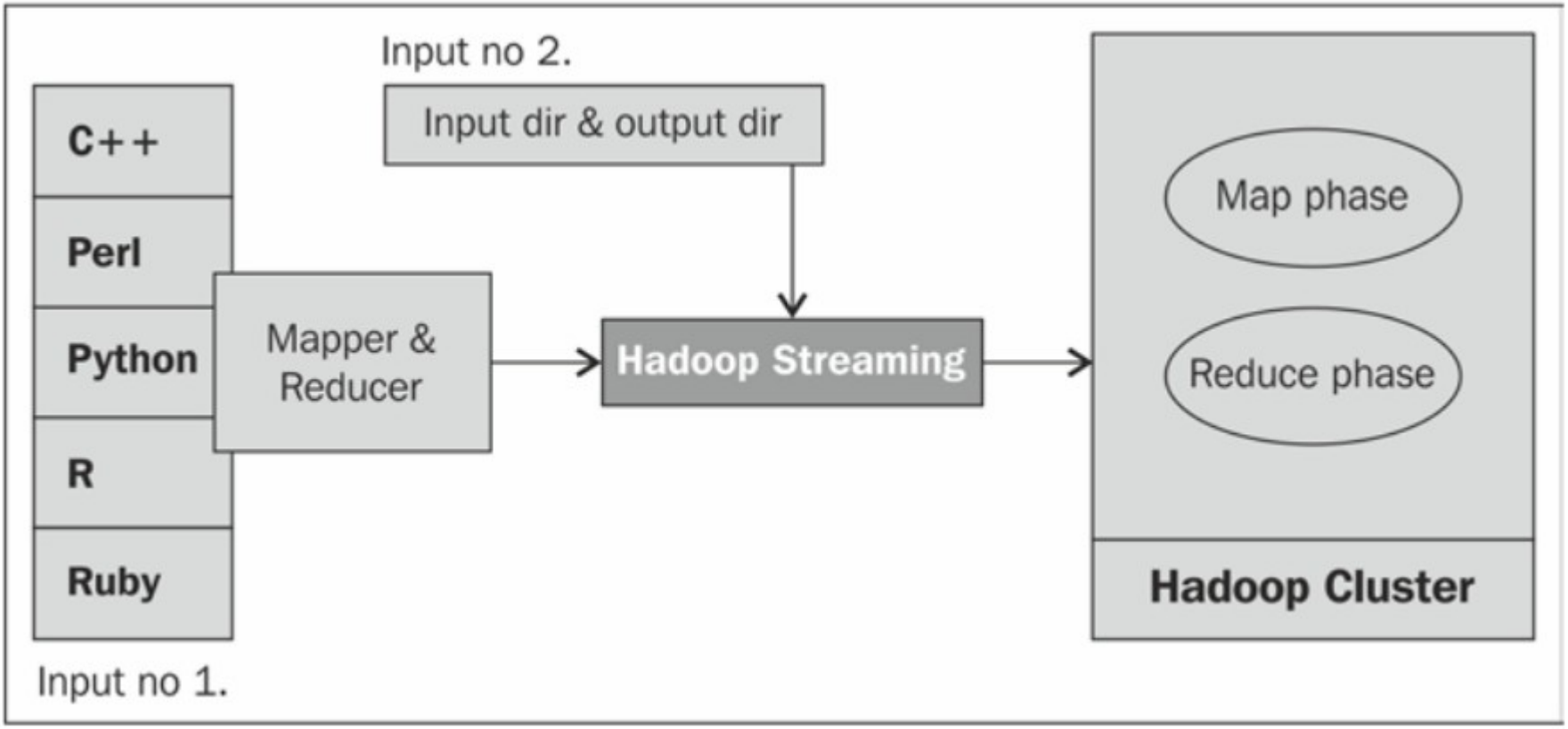
- The input reader contains the complete logic about the data it is reading.
- Suppose we want to read an image then we have to specify the logic in the input reader so that it can read that image data and finally it will generate key-value pairs for that image data.

Hadoop Streaming

- If we are reading an image data then we can generate key-value pair for each pixel where the key will be the location of the pixel and the value will be its color value from (0-255) for a colored image.
- Now this list of key-value pairs is fed to the Map phase and Mapper will work on each of these key-value pair of each pixel and generate some intermediate key-value pairs which are then fed to the Reducer after doing shuffling and sorting then the final output produced by the reducer will be written to the HDFS.

miTu skillologies
tusharkute.com

Hadoop Streaming



Example:

- ```
  hadoop jar
  /usr/local/hadoop/share/hadoop/tools/li
  b/hadoop-streaming-3.2.3.jar
  -file mapper1.py
  -mapper mapper1.py
  -file reducer1.py
  -reducer reducer1.py
  -input /input1/fruits.txt
  -output /output11
  ```
  ```

Hadoop Streaming Command

```
${HADOOP_HOME}/bin/hadoop \
        jar $HADOOP_HOME/contrib/*.jar \                    Line 1
        -input /app/haadoop/input \                         Line 2
        -output /app/haadoop/output \                       Line 3
        -file /usr/local/hadoop/code_mapper.R \             Line 4
        -mapper code_mapper.R \                             Line 5
        -file /usr/local/hadoop/code_reducer.R \            Line 6
        -reducer code_reducer.R                             Line 7
```

tusharkute.com

Streaming Command Options

Option	Description
-input directory_name or filename	Input location for the mapper.
-output directory_name	Input location for the reducer.
-mapper executable or JavaClassName	The command to be run as the mapper
-reducer executable or script or JavaClassName	The command to be run as the reducer
-file file-name	Make the mapper, reducer, or combiner executable available locally on the compute nodes
-inputformat JavaClassName	By default, TextInputformat is used to return the key-value pair of Text class. We can specify our class but that should also return a key-value pair.

tusharkute.com

Streaming Command Options

-outputformat JavaClassName	By default, TextOutputformat is used to take key-value pairs of Text class. We can specify our class but that should also take a key-value pair.
-partitioner JavaClassName	The Class that determines which key to reduce.
-combiner streamingCommand or JavaClassName	The Combiner executable for map output
-verbose	The Verbose output.
-numReduceTasks	It Specifies the number of reducers.
-mapdebug	Script to call when map task fails
-reducedebug	Script to call when reduce task fails

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com