Hadoop

MapReduce

Objectives

- What is MapReduce?
- MapReduce and Yarn
- MapReduce with Shell Scripts
- MapReduce with Python
- MapReduce with Java
- Labs

What is MapReduce?

- From the developer's perspective, MapReduce is two things:
  - Map, A Program / Function / Class that produces a set of key-value pairs.
  - Reduce, A Program / Function / Class that accepts the key-value pairs produced from Map. Reduce takes this set of key-value pairs and produces the desired information.
  - As we have our data files split across potentially many servers, Map programs are pushed to the servers and they operate on their "piece" of data producing their portion.
  - Reduce collects these pieces of data, combines the data, and produces a final product.

MapReduce and Yarn

- In V1 of Hadoop, MapReduce had the responsibility of cluster management and resource allocation. This is in addition to its responsibility to process data requests.

- In V2 of Hadoop, Yarn was introduced to remove the cluster management responsibilites of MapReduce. So, MapReduce now has no cluster management responsibilities. MapReduce processes data.

MapReduce With Shell Scripts

- Shell Scripts provide the easiest mechanism for accessing the streaming tools of UNIX. However, the shell is somewhat lacking in advanced programming constructs and access to native libraries.

- But, for some data extraction and processing applications, it may be all you need.

- Any UNIX tool (or any programming tool) used to write or read data can easily be stored in a shell script to produce Map-type output and read the output for a Reduce-type result.

- Also, Shell Scripts are a nice tool for wrapping other technologies for launching MapReduce, Pig, Hive, Sqoop, and Spark jobs.

MapReduce With Python

- As Python is a great tool for working with data files, programmers may opt to use this language for their Hadoop processing.
- Python has no native MapReduce functionality but can easily write to a UNIX stream.
- Hadoop provides a series of java classes to convert streams of data to MapReduce classes. These classes are stored in a jar file named hadoop-streaming*.jar.
- We will write a python Map script to generate the data which we'll write to stdout. We will write a python Reduce script which will read from stdin and write the final results to stdout.

MapReduce With Java

- Java is the native language of Hadoop. Many of the Hadoop tools are expecting Java classes to represent the Map and Reduce phases of a Hadoop job along with intermediary classes for combining data, sorting data, etc.

- The Hadoop API provides every piece of functionality available in the Hadoop environment.

- Java Mapping and Reducing is not an easy task for non-Java programmers (it's actually not all that easy for Java programmers). However, Java provides the most granular control over the functioning of Mapping and Reducing. Java programs have complete control over all aspects of the Hadoop framework.

Labs

- Emulate a MapReduce Program with a simple shell script.
- Use our shell scripting with other unix tools to provide a true MapReduce program.
- Emulate a MapReduce Program with a Python Map script and a Python Reduce script.
- Wrap our Python scripts in a shell script to provide a true MapReduce program.
- Compile native Java programs and submit to Hadoop as a native Hadoop application.