Recitation 1

CS435: Introduction to Big Data

GTA: Bibek R. Shrestha
Email: cs435@cs.colostate.edu

Jan 19, 2018
Today…

- We will go through introductory slides.
- We will see how to configure Hadoop in standalone and cluster mode. (one more mode - pseudo-distributed mode)
- We will run word-count example in both of the Hadoop modes.
- We will not use HDFS for input or output.
- We will not go into the actual java code.
Score Distribution

There will be 3 PAs for this course. Each PA carries 10 points. PA0 & PA1 constitute 1st assignment and thus 10 points (3 + 7).

For PA0:

- 1 point for Initializing Hadoop cluster
- 1 point for Running basic HDFS operations
- 1 point for Running MapReduce example
Types of Hadoop Daemons

- **Namenode**: master node manages the filesystem namespace, communicate with other daemons and make decisions
- **Secondary Namenode**: backup namenode
- **Jobtracker**: keeps tracks of all running jobs in the cluster
- **Datanode**: worker node that actually run map/reduce task on HDFS blocks
- **Tasktracker**: maintains task (map, reduce, shuffle) at each datanode and communicate with Jobtracker
Understanding Map Reduce
Data Flow in Hadoop
WordCount Example

The Overall MapReduce Word Count Process

Input: K1,V1

Splitting:
- Deer Bear River
- Car Car River
- Deer Car Bear

Mapping:
- List(K2,V2)
  - Deer, 1
  - Bear, 1
  - River, 1

Shuffling:
- K2,List(V2)
  - Bear, (1,1)

Reducing:
- Bear, 2

Final Result:
- List(K3,V3)
  - Bear, 2
  - Car, 3
  - Deer, 2
  - River, 2
Notes

- Use unique nodes and ports while configuring Hadoop cluster.
- Always make sure the output folder you have specified when running your job in Hadoop does not exist.
- Do not format your namenode. It will delete all the files in HDFS. (format namenode during first time configuration only or if absolutely needed.)
- By default, each node can have two map tasks and two reduce task, where each task can utilize ~1GB of memory. Thus, make sure your node can use 4GB memory at any time.
- The number of map task by default is the number of input splits which by default is 64MB (size of HDFS block). The number of reduce task executed depends on many factors (number of data blocks, slots available in datanodes, etc). Using multiple reduce tasks is essential when the single reducer is the bottleneck for the system.
In the next recitation,

- We will use HDFS for storing input and output for the MapReduce job.
- We will look into the word-count MapReduce program provided.