Word Count Using Hadoop

Steps to Run the WordCount Job

1. **Start the Hadoop services:**

   start-dfs.sh

   start-yarn.sh

2. **Create input directory in HDFS:**

   hdfs dfs -mkdir -p /user/<your-username>/wordcount/input

3. **Copy `input.txt` to HDFS:**

   hdfs dfs -put input.txt /user/sai/wordcount/input

hdfs dfs -put input.txt /user/sai/wordcount/inputz

4. **Run the WordCount job:**

   yarn jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar wordcount /user/sai/wordcount/input /user/<your-username>/wordcount/output

5. **View the output:**

   hdfs dfs -cat /user/sai/wordcount/output/part-r-00000

Edit the `input.txt` file locally:

You can use any text editor (like `nano`, `vim`, or `gedit`) to open and edit the file.

For example, using `nano`:

nano input.txt

Make your changes (e.g., add more words) and save the file.

Overwrite the existing file in HDFS:

After editing the file locally, you'll need to update the file in HDFS. First, remove the old file from HDFS:

hdfs dfs -rm /user/sai/wordcount/input/input.txt

Then, copy the updated `input.txt` file to HDFS:

hdfs dfs -put input.txt /user/sai/wordcount/input/

Re-run the WordCount job:

After updating the input file, you can re-run the WordCount job to see the results with the new content.

yarn jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar wordcount /user/sai/wordcount/input /user/sai/wordcount/output

View the new output:

Finally, check the output:

hdfs dfs -cat /user/sai/wordcount/output/part-r-00000