Water Quality

Data Ingest

```
[jc8017@login-2-1 data_injest]$ bash run_water_injest.bash
put: `/user/ma4759/infrastructure/water_quality/inputdata.csv': File exists
```

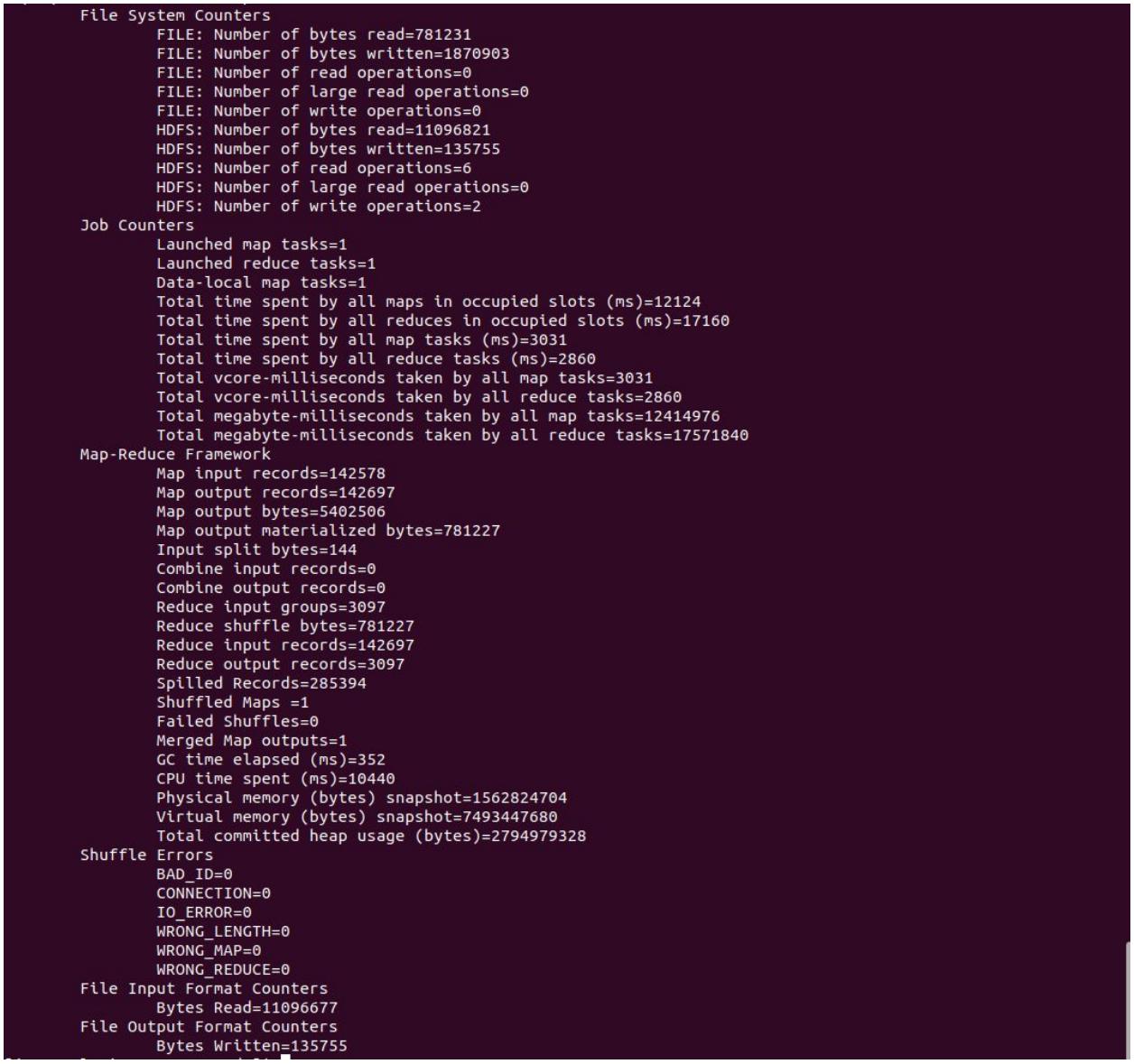
Clean Data (ETL Code)

```
[jc8017@login-2-1 etl_code]$ bash run_water_clean.bash
20/11/28 16:03:10 INFO fs.TrashPolicyDefault: Moved: 'hdfs://dumbo/user/ma4759/infrastructure/water_quality/clean_out
put' to trash at: hdfs://dumbo/user/jc8017/.Trash/Current/user/ma4759/infrastructure/water_quality/clean_output
Note: Clean.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
added manifest
adding: CleanMapper.class(in = 2114) (out= 863)(deflated 59%)
adding: CleanReducer.class(in = 1559) (out= 644)(deflated 58%)
adding: Clean.class(in = 1389) (out= 808)(deflated 41%)
20/11/28 16:03:15 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
 Tool interface and execute your application with ToolRunner to remedy this.
20/11/28 16:03:15 INFO input.FileInputFormat: Total input paths to process : 1
20/11/28 16:03:16 INFO mapreduce.JobSubmitter: number of splits:1
20/11/28 16:03:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1604346392376_5213
20/11/28 16:03:16 INFO impl.YarnClientImpl: Submitted application application_1604346392376_5213
20/11/28 16:03:16 INFO mapreduce.Job: The url to track the job: http://babar.es.its.nyu.edu:8088/proxy/application_16
04346392376_5213/
20/11/28 16:03:16 INFO mapreduce.Job: Running job: job_1604346392376_5213
20/11/28 16:03:20 INFO mapreduce.Job: Job job_1604346392376_5213 running in uber mode : false
20/11/28 16:03:20 INFO mapreduce.Job:  map 0% reduce 0%
20/11/28 16:03:25 INFO mapreduce.Job:  map 100% reduce 0%
20/11/28 16:03:32 INFO mapreduce.Job:  map 100% reduce 100%
20/11/28 16:03:34 INFO mapreduce.Job: Job job_1604346392376_5213 completed successfully
20/11/28 16:03:34 INFO mapreduce.Job: Counters: 49
```

```
        File System Counters
                FILE: Number of bytes read=4785623
                FILE: Number of bytes written=9879649
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=20152907
                HDFS: Number of bytes written=11096677
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=11988
                Total time spent by all reduces in occupied slots (ms)=16038
                Total time spent by all map tasks (ms)=2997
                Total time spent by all reduce tasks (ms)=2673
                Total vcore-milliseconds taken by all map tasks=2997
                Total vcore-milliseconds taken by all reduce tasks=2673
                Total megabyte-milliseconds taken by all map tasks=12275712
                Total megabyte-milliseconds taken by all reduce tasks=16422912
        Map-Reduce Framework
                Map input records=142708
                Map output records=142704
                Map output bytes=11534646
                Map output materialized bytes=4785619
                Input split bytes=132
                Combine input records=0
                Combine output records=0
                Reduce input groups=142578
                Reduce shuffle bytes=4785619
                Reduce input records=142704
                Reduce output records=142578
                Spilled Records=285408
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=332
                CPU time spent (ms)=10530
                Physical memory (bytes) snapshot=1551941632
                Virtual memory (bytes) snapshot=7480836096
                Total committed heap usage (bytes)=2788163584
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=20152775
        File Output Format Counters
                Bytes Written=11096677
```

Analyze Code (Ana_code)

```
[jc8017@login-2-1 ana_code]$ bash run_water_analyze.bash
20/11/28 16:15:46 INFO fs.TrashPolicyDefault: Moved: 'hdfs://dumbo/user/ma4759/infrastructure/water_quality/analysis_
output' to trash at: hdfs://dumbo/user/jc8017/.Trash/Current/user/ma4759/infrastructure/water_quality/analysis_output
1606598146077
Note: AnalyzeWater.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
added manifest
adding: AnalyzeWaterMapper.class(in = 3279) (out= 1482)(deflated 54%)
adding: AnalyzeWaterReducer.class(in = 2079) (out= 931)(deflated 55%)
adding: AnalyzeWater.class(in = 1407) (out= 812)(deflated 42%)
adding: AnalyzeWritable.class(in = 2529) (out= 1102)(deflated 56%)
20/11/28 16:15:51 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
 Tool interface and execute your application with ToolRunner to remedy this.
20/11/28 16:15:52 INFO input.FileInputFormat: Total input paths to process : 1
20/11/28 16:15:52 INFO mapreduce.JobSubmitter: number of splits:1
20/11/28 16:15:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1604346392376_5220
20/11/28 16:15:52 INFO impl.YarnClientImpl: Submitted application application_1604346392376_5220
20/11/28 16:15:53 INFO mapreduce.Job: The url to track the job: http://babar.es.its.nyu.edu:8088/proxy/application_16
04346392376_5220/
20/11/28 16:15:53 INFO mapreduce.Job: Running job: job_1604346392376_5220
20/11/28 16:15:57 INFO mapreduce.Job: Job job_1604346392376_5220 running in uber mode : false
20/11/28 16:15:57 INFO mapreduce.Job:  map 0% reduce 0%
20/11/28 16:16:02 INFO mapreduce.Job:  map 100% reduce 0%
20/11/28 16:16:08 INFO mapreduce.Job:  map 100% reduce 100%
20/11/28 16:16:11 INFO mapreduce.Job: Job job_1604346392376_5220 completed successfully
20/11/28 16:16:11 INFO mapreduce.Job: Counters: 49
```



Data Profiling:

```
[jc8017@login-2-1 profiling_code]$ bash run_water_profile.bash
rm: `/user/ma4759/infrastructure/water_quality/profiling_output': No such file or directory
Note: CountCounty.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
added manifest
adding: CountCountyMapper.class(in = 1455) (out= 558)(deflated 61%)
adding: CountCountyReducer.class(in = 1671) (out= 708)(deflated 57%)
adding: CountCounty.class(in = 1419) (out= 813)(deflated 42%)
20/11/28 16:23:51 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the
 Tool interface and execute your application with ToolRunner to remedy this.
20/11/28 16:23:51 INFO input.FileInputFormat: Total input paths to process : 1
20/11/28 16:23:52 INFO mapreduce.JobSubmitter: number of splits:1
20/11/28 16:23:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1604346392376_5225
20/11/28 16:23:52 INFO impl.YarnClientImpl: Submitted application application_1604346392376_5225
20/11/28 16:23:52 INFO mapreduce.Job: The url to track the job: http://babar.es.its.nyu.edu:8088/proxy/application_16
04346392376_5225/
20/11/28 16:23:52 INFO mapreduce.Job: Running job: job_1604346392376_5225
20/11/28 16:23:58 INFO mapreduce.Job: Job job_1604346392376_5225 running in uber mode : false
20/11/28 16:23:58 INFO mapreduce.Job:  map 0% reduce 0%
20/11/28 16:24:02 INFO mapreduce.Job:  map 100% reduce 0%
20/11/28 16:24:07 INFO mapreduce.Job:  map 100% reduce 100%
20/11/28 16:24:08 INFO mapreduce.Job: Job job_1604346392376_5225 completed successfully
20/11/28 16:24:08 INFO mapreduce.Job: Counters: 49
```

```
        File System Counters
                FILE: Number of bytes read=1644
                FILE: Number of bytes written=311765
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=135902
                HDFS: Number of bytes written=23
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=8508
                Total time spent by all reduces in occupied slots (ms)=15258
                Total time spent by all map tasks (ms)=2127
                Total time spent by all reduce tasks (ms)=2543
                Total vcore-milliseconds taken by all map tasks=2127
                Total vcore-milliseconds taken by all reduce tasks=2543
                Total megabyte-milliseconds taken by all map tasks=8712192
                Total megabyte-milliseconds taken by all reduce tasks=15624192
        Map-Reduce Framework
                Map input records=3097
                Map output records=3097
                Map output bytes=27873
                Map output materialized bytes=1640
                Input split bytes=147
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=1640
                Reduce input records=3097
                Reduce output records=1
                Spilled Records=6194
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=91
                CPU time spent (ms)=2060
                Physical memory (bytes) snapshot=913694720
                Virtual memory (bytes) snapshot=7471521792
                Total committed heap usage (bytes)=2413297664
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=135755
        File Output Format Counters
                Bytes Written=23
```