Oden Take Home Exercise

Apache Flink Solution:

1. Used Apache Flink with local installation run directly from Scala code (build with SBT)

2. Used streaming context and read epa-http.txt directly into DataStream[String] and mapped to DataStream[WebLog] where WebLog is a case class for each data entry including IP Address, timestamp and number of bytes. A rowtime is added to each timestamp entry for windowing.

3. Aggregation and windowing are performed with Flink table methods equivalent to SQL – group by IP address and window and sum on number of bytes. (For entire dataset, window is set to 25 hours to capture all data)

4. Results are generated by generating DataStream[Results] and DataStream[WindowedResults] from table where Results and WindowedResults are case classes.

5. To enable output formatting (especially with DateTime), print to STDOUT and write to file are done from Scala code directly within map function.

6. Results for complete file in epa-results.txt and for tumbling window in windowed-epa-results.txt.

7. Can run with SBT run from command line or from SBT shell. (May need to edit filepath parameter for epa-http.txt)

Apache Spark Solution:

1. Used Apache Spark with stand alone local with a one node cluster run directly from Scala code as SparkLauncher. Actual Scala Spark app is built as a Fat Jar and then run from the SparkLauncher.

2. Use Gradle (rather than SBT) as Fat Jar plugin is preferable to SBT assembly plugin (personal ease of use at present)

3. Read epa-http.txt directly as String DataFrame and transform through WebBlog case class to DataFrame with schema.

4. DataFrame methods enable aggregation and windowing

5. To enable output formatting, write to file are done from Scala code directly within map function after dataframe has been collected to a Scala Array[Row].

6. Results for complete file in epa-results.txt and for tumbling window in windowed-epa-results.txt.

7. Need to build application “Oden” with “./gradlew build” from command line first. Then run “OdenLauncher” from command line with “./gradlew run”. (May need to edit filepath parameter for epa-http.txt in application “Oden” and edit path in setAppResource within “OdenLauncher”)