CS 378 – Big Data Programming

Lecture 11

AVRO Formats

Review

- Assignment 5 – Avro Objects

- Questions/Issues?
  - `datum()`

- Using the latest pom.xml
  - The executable Java class is **not** defined in this pom.xml
  - You must specify the fully qualified Java class name as the first argument to your app, or
  - Add the executable class definition back into pom.xml

AVRO File Formats

- `TextOutputFormat`
  - How are various key and value types handled?
  - Recall that `TextOutputFormat` will cause `toString()` to be called

- `AvroKey<CharSequence>`
  - Acts like `Text`, so it just returns its string value

- `AvroValue<WordStatisticsData>`
  - Returns the value created by the `toString()` method

AVRO File Formats

- `TextOutputFormat`

- `AvroKey<Pair<CharSequence, WordStatisticsData>>`
  - Usually used as the key to the `write()` method, with value `NullWritable`
  - Generates a string representation in the `toString()` method of `Pair`
  - In this form: `{“key”:` *theKey*`, “value”:` *theValue* `}`
  - *theKey* comes from `CharSequence`, so just a string
  - *theValue* comes from `WordStatisticsData`, so an Avro text representation is generated
  - `{ “document_count”: ….. }`

AVRO File Formats

- `AvroKeyValueOutputFormat`
  - Creates a generic Avro record with a "key" field and a "value" field
    - Like what we saw with `AvroKey<Pair< K, V >>`
  - Avro container file (binary)
  - Can be read in using: `AvroKeyValueInputFormat`

AVRO File Formats

- `AvroKeyOutputFormat<T>`
  - Extends
  - `AvroOutputFormatBase(AvroKey<T>, NullWritable>)`
  - Only the key is output, value is ignored
  - Avro container file (binary format)

  - Can be read in using: `AvroKeyInputFormat`

AVRO File Formats

- `AvroSequenceFileOutputFormat`
  - Sequence file output format that can handle `AvroKey` and `AvroValue` in addition to `Writable`
  - Can be read with: `AvroSequenceFileInputFormat`

AVRO File Formats

- `AvroKeyValueInputFormat`
  - Reads generic Avro records with a “key” field and a “value” field
  - Avro container file (binary)
  - Data should have been written with: `AvroKeyValueOutputFormat`

AVRO File Formats

- AvroKeyInputFormat
  - Extends
  - `FileInputFormat(AvroKey<T>, NullWritable>)`
  - Only the key is read, value is ignored
  - Reads a Avro container file (binary format)
  - Data should have been written with: `AvroKeyOutputFormat`

AVRO File Formats

- AvroSequenceFileInputFormat
  - Input format that can read sequence files that support Avro types
  - Data should have been written with: `AvroSequenceFileOutputFormat`

Design Pattern

- Structured to hierarchical design pattern

- Data sources linked by some foreign key
- Data is structured and row based
  - For example, from databases
- Data is semi-structured and event based
  - Web logs

Design Pattern

- Structured to hierarchical design pattern

- MultipleInputs
  - Able to accept data inputs from different formats
  - Mappers load and parse the input into a cohesive format
  - Prepared for work in the reducer
  - Map output key will be the unifying element of the hierarchical record

- Combiners don't help, as they don't "reduce" the data (make it smaller)

Design Pattern

- Structured to hierarchical design pattern

- Reducer takes all the data associated with a key
- Builds the structure to be output

- Example:
  - User session contains info about the user (IP, browser, …)
  - An array of actions (page views, clicks, …)

MapReduce in Hadoop

Figure 2.4, Hadoop - The Definitive Guide

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Sessionizing Web Logs

- Create user sessions from individual web log entries

- Represents all the actions by a user
- Allows later analysis to “replay” the user actions

- Collect measures and metrics about user behavior
  - Pages viewed, time on page, clicks
  - Path through the site, entry to the site (from a search engine?)

Assignment 5

- Bootstrap script (control classpath order)
- pom.xml provided
  - Use this one, as AVRO with Hadoop is version sensitive
  - Select AMI version 2.4.7 when defining your cluster
- Examples of WordCount provided
- Implement an AVRO object for WordStatistics data
  - Call it `WordStatisticsData`
  - Mapper output:
    - `Text, AvroValue<WordStatisticsData>`
  - Reducer output:
    - `AvroKey<Pair<CharSequence, WordStatisticsData>>`
  - Output file format: `TextOutputFormat` (like WordCountD)