CS 686: Special Topics in Big Data

MapReduce Tips

Lecture 21

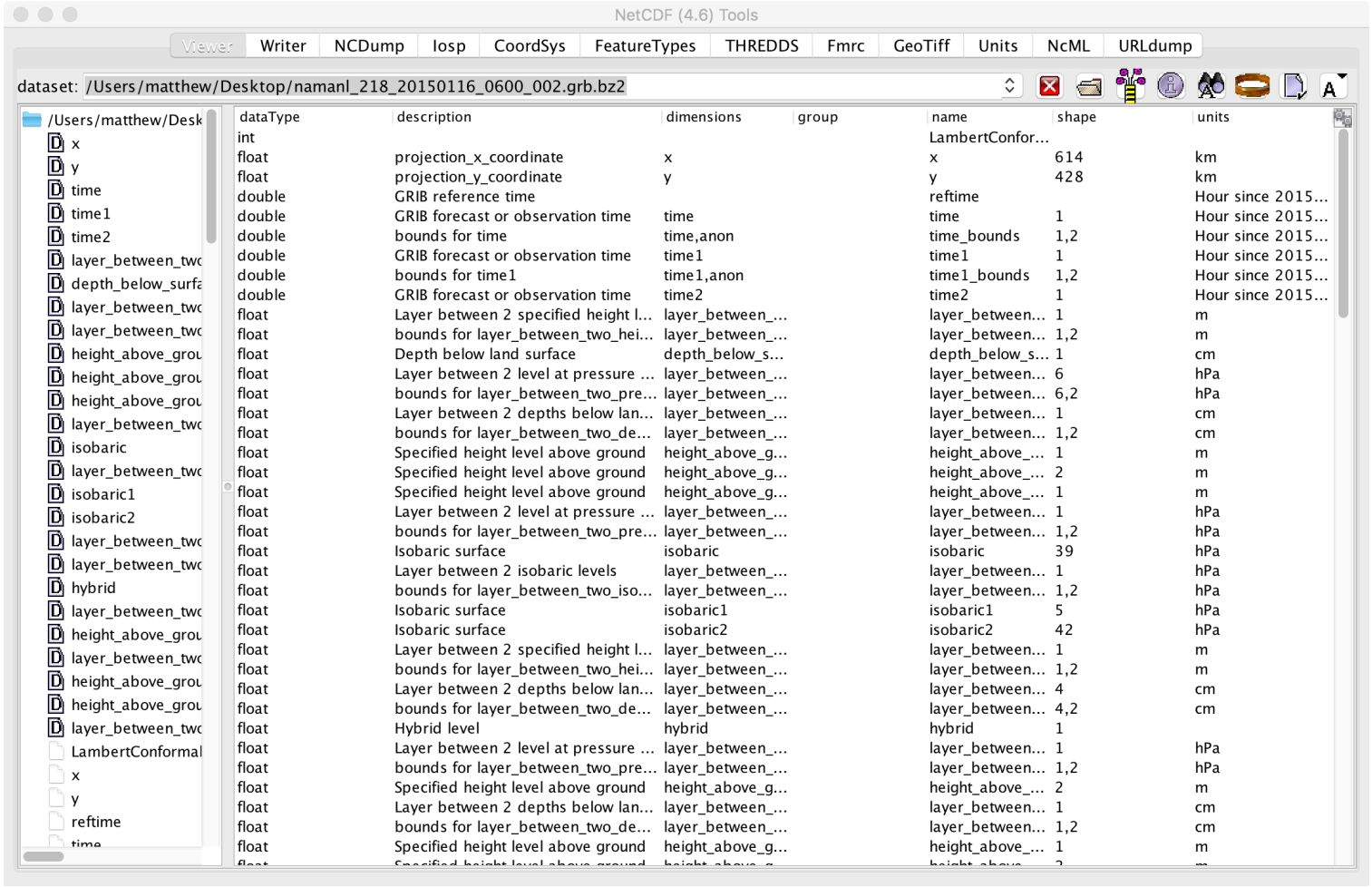
Cluster Updates

- There seems to have been a small problem with how YARN was configured to schedule jobs
  - Aggressively started reduce tasks before the mappers were all finished
- This may have led to some jobs hanging
- We'll keep monitoring the situation, but if something is out of the ordinary definitely let me know!

P2 Updates

- The P2 spec has been updated to clarify some minor points
- I also added an original (GRIB) file from the NAM that you can download and play with
  - Read these files with the Java NetCDF library (also linked from the dataset page)
- Particularly useful: the toolsUI
- Also useful: a list of all possible Geohashes

NetCDF Tools



dataType	description	dimensions	group	name	shape	units
int				LambertConfor...		
float	projection_x_coordinate	x		x	614	km
float	projection_y_coordinate	y		y	428	km
double	GRIB reference time			reftime		Hour since 2015...
double	GRIB forecast or observation time	time		time	1	Hour since 2015...
double	bounds for time	time,anon		time_bounds	1,2	Hour since 2015...
double	GRIB forecast or observation time	time1		time1	1	Hour since 2015...
double	bounds for time1	time1,anon		time1_bounds	1,2	Hour since 2015...
double	GRIB forecast or observation time	time2		time2	1	Hour since 2015...
float	Layer between 2 specified height l...	layer_between_...		layer_between...	1	m
float	bounds for layer_between_two_hei...	layer_between_...		layer_between...	1,2	m
float	Depth below land surface	depth_below_s...		depth_below_s...	1	cm
float	Layer between 2 level at pressure ...	layer_between_...		layer_between...	6	hPa
float	bounds for layer_between_two_pre...	layer_between_...		layer_between...	6,2	hPa
float	Layer between 2 depths below lan...	layer_between_...		layer_between...	1	cm
float	bounds for layer_between_two_de...	layer_between_...		layer_between...	1,2	cm
float	Specified height level above ground	height_above_g...		height_above_...	1	m
float	Specified height level above ground	height_above_g...		height_above_...	2	m
float	Specified height level above ground	height_above_g...		height_above_...	1	m
float	Layer between 2 level at pressure ...	layer_between_...		layer_between...	1	hPa
float	bounds for layer_between_two_pre...	layer_between_...		layer_between...	1,2	hPa
float	Isobaric surface	isobaric		isobaric	39	hPa
float	Layer between 2 isobaric levels	layer_between_...		layer_between...	1	hPa
float	bounds for layer_between_two_iso...	layer_between_...		layer_between...	1,2	hPa
float	Isobaric surface	isobaric1		isobaric1	5	hPa
float	Isobaric surface	isobaric2		isobaric2	42	hPa
float	Layer between 2 specified height l...	layer_between_...		layer_between...	1	m
float	bounds for layer_between_two_hei...	layer_between_...		layer_between...	1,2	m
float	Layer between 2 depths below lan...	layer_between_...		layer_between...	4	cm
float	bounds for layer_between_two_de...	layer_between_...		layer_between...	4,2	cm
float	Hybrid level	hybrid		hybrid	1	
float	Layer between 2 level at pressure ...	layer_between_...		layer_between...	1	hPa
float	bounds for layer_between_two_pre...	layer_between_...		layer_between...	1,2	hPa
float	Specified height level above ground	height_above_g...		height_above_...	2	m
float	Layer between 2 depths below lan...	layer_between_...		layer_between...	1	cm
float	bounds for layer_between_two_de...	layer_between_...		layer_between...	1,2	cm
float	Specified height level above ground	height_above_g...		height_above_...	1	m

Reading the Dataset

- You can add a directory, file, or a pattern to read for your MapReduce jobs
- Don't just give the NAM directory; it has both the full dataset and the mini dataset
  - Not too big of a deal, but duplicates some information
- Instead, specify a pattern:
  - `/tmp/cs686/nam/nam_2015\*`
  - The wildcard escape is needed if providing the path from your terminal

Testing Your Jobs

- It's a good idea to use the mini dataset for testing
  - This can still take a bit of time to run
- Another recommendation: create an even smaller dataset for rapid development
  - ```
    hdfs dfs -cat \
      /tmp/cs686/nam/nam_mini.tdv \
      | head -n 100 | shuf > nam_tiny.tdv
    ```
- Then run your job on just one of the bass nodes (don't submit the job on bass01)
    ```

Operating on Local Files

- You can specify file:///home4/username/file as an input or output to use non-HDFS paths

Cleaning Up

- One thing to remember: hitting Ctrl+C isn't going to kill your job
- Be sure to:
  - `yarn application -kill <app_id>`
- Test your applications with the mini dataset before running across the entire dataset!

Being Lazy

You can use the LazyOutputFormat to avoid writing empty files during the reduce phase

```
import org.apache.hadoop.mapreduce.lib.output.LazyOutputFormat;

. . .

LazyOutputFormat.setOutputFormatClass(job, TextOutputFormat.class);
```

Cleanup() Method

Let's assume you populate a HashMap with values during the reduce phase. You can then emit a condensed version during cleanup:

```
@Override
protected void cleanup(Context context)
throws IOException, InterruptedException {
    for (Text geohash : hottest.keySet()) {
        Double temp = hottest.get(geohash);
        context.write(geohash, new DoubleWritable(temp));
    }
}
```

Custom Writables

- Many of the questions want to know more than one thing
  - For example, both **when** and **where** something happened
- You can emit text separated by tabs (or whatever character you like most)
- **Or** you can create your own WritableComparable
  - Best practice, but not required

Custom Output Formats

- You can also write your own output formats
  - Not **too** much work – implement some methods
- Not required for the assignment, but definitely go for it if you feel it helps!
- Here's how you can write your own format that doesn't produce empty files:
  - http://whiteycode.blogspot.it/2012/06/hadoop-removing-empty-output-files.html