CS 686: Special Topics in Big Data

NAM Analysis

Lecture 22

Cluster News

- ACLs should be set up
    - Any volunteers want to test?
- Limit: 5 concurrent jobs
    - FIFO
    - Submit your job, wait in the queue…
- New ground rule: during labs, only operate on nam_mini.tdv
    - Even better: copy it to your home directory and run jobs locally (submit on any machine other than bass01)

Grading

- Q: How will we grade P2?
- A: For most of the questions, we'll look at the answers you got
  - For a couple, I'll give you a new, unseen dataset to run your MapReduce applications on
    - Also from the NAM, .tdv file, but very small
  - We'll walk through your logic for some of the jobs
- A note: if you come up with a unique way of tacking one of the problems, I'm even happier!

Grading pt. 2

- Anybody done with all the warm ups?
- Anybody done with Deliverable I?
- I'll come by to grade

Tippity Top

- Q: What do you mean, "top 3" ?!
  (It's a bit vague…)
- A: Well, it depends. It could mean the highest accumulated values (total snow). It could also mean the average (average snow over the year).
  - Maybe your analysis only finds one point – in that case, it's okay
  - Maybe you find 10,000 (this is why I put a limit on what you report)
- For example…

Snow Depth

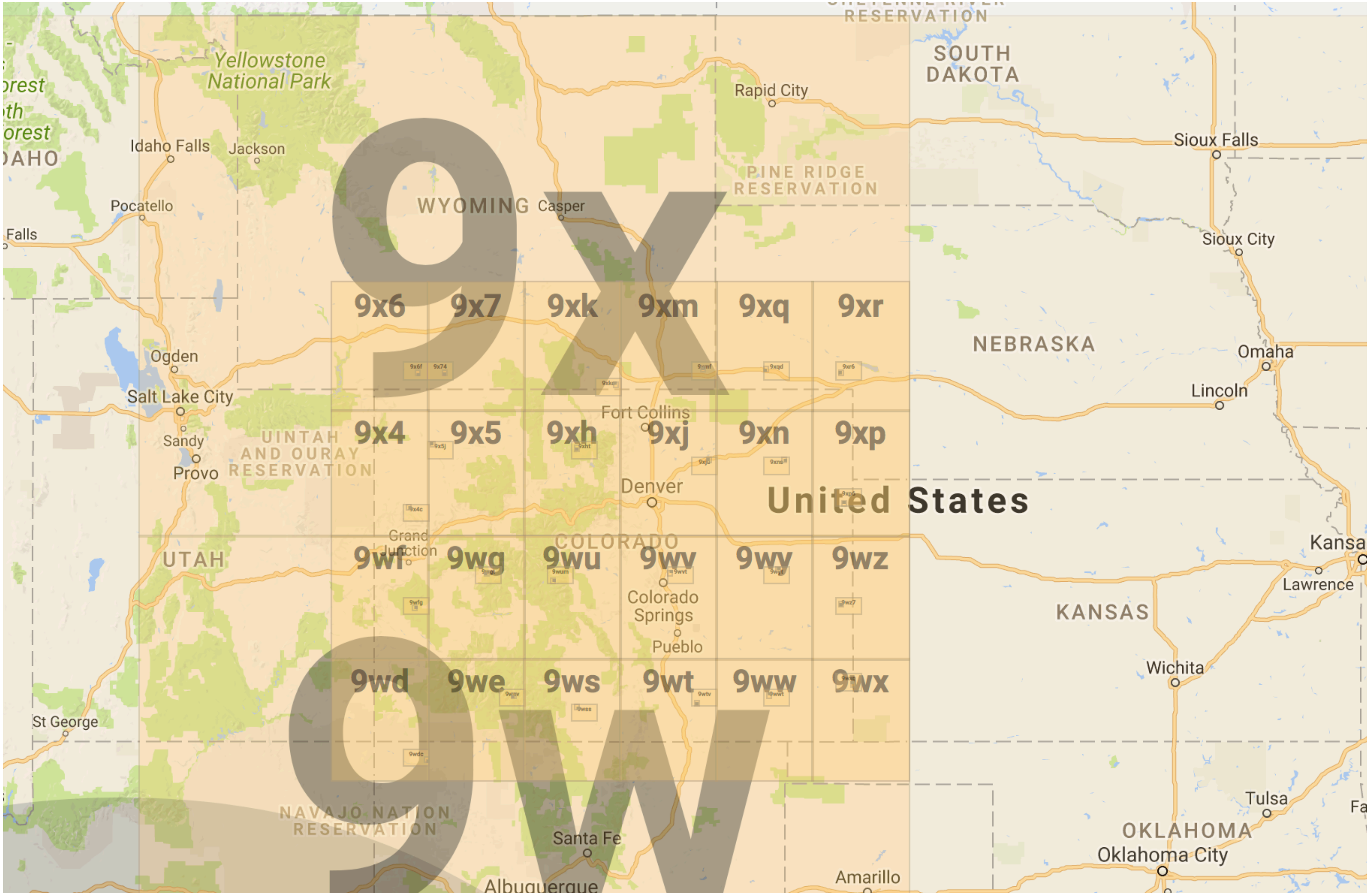
- Let’s look at the snow depth problem. Perhaps you will emit (*these are just **ideas**! ☺ *):
  - The snow depth itself
  - 0 if no snow, 1 if there is
  - The time the depth was greater than 0
- Then report, for any geohashes that did **not** have a
  - The average snow depth, sort, and find the highest
  - The percentage of time there is snow at each location
  - The number of months when snow was present

Defining Regions via Geohash

- One question asks about the bay area. How do we define this?
- My recommendation:

  http://geohash.gofreerange.com
- Visually locate the areas you are interested and note their Geohashes in a list
- Then filter based on the entries in the list

Defining Colorado



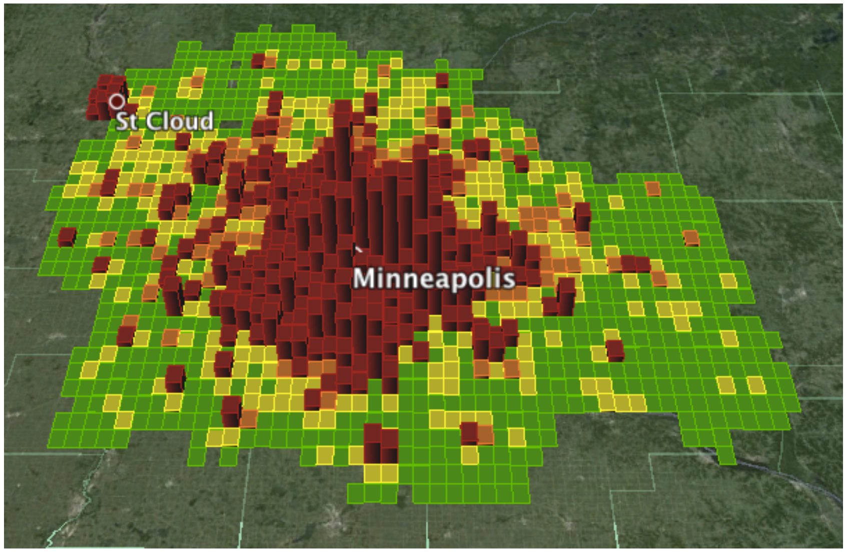
Constraining our Analysis

- For a few questions, I ask for a specific Geohash precision
  - For example, four-character Geohashes
- To do this, just chop the extra characters off the string:
  - 9xjq94b → 9xjq
- Use the Geohash as a key from your Mapper
- Reducer gets all the data points that fell within 9xjq!

Interesting: geohash2kml

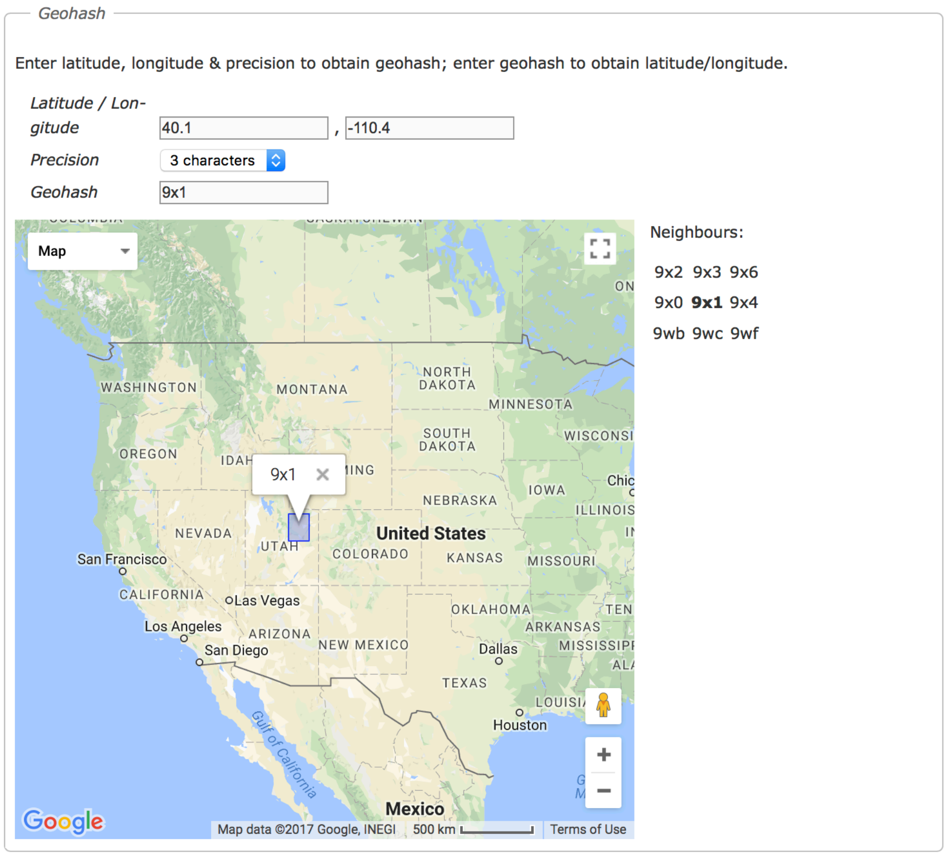
Here's a library for generating Google Earth visualizations:

https://github.com/abeusher/geohash2kml



Also Interesting: Geohash + Map

http://www.movable-type.co.uk/scripts/geohash.html



Geohash

Enter latitude, longitude & precision to obtain geohash; enter geohash to obtain latitude/longitude.

Latitude / Longitude 40.1 , -110.4

Precision 3 characters

Geohash 9x1

Neighbours:

9x2	9x3	9x6
9x0	**9x1**	9x4
9wb	9wc	9wf

Sanity Checking

- If you have a very small subset, you can open it in a spreadsheet application
  - Separate the columns by tab characters “\t”
- I prefer to use **awk**:

```
awk -F'\t' '{print $14, $51}' nam_mini.tdv | head -n 5
# (Prints the first five values for features 14 and 51)

hdfs dfs -cat /some/files/somewhere/part-r-\* \
    | awk -F'\t' '{print $2, $1}' | sort -n
# Swaps the positions of columns 1 and 2, and then sorts
  numerically. Good for manipulating MR outputs.
```