GIS 2: Rabin Fingerprinting

Content Defined Chunking

CDC algorithms such as Rabin Fingerprints let you chunk a file based on the content of the file itself.

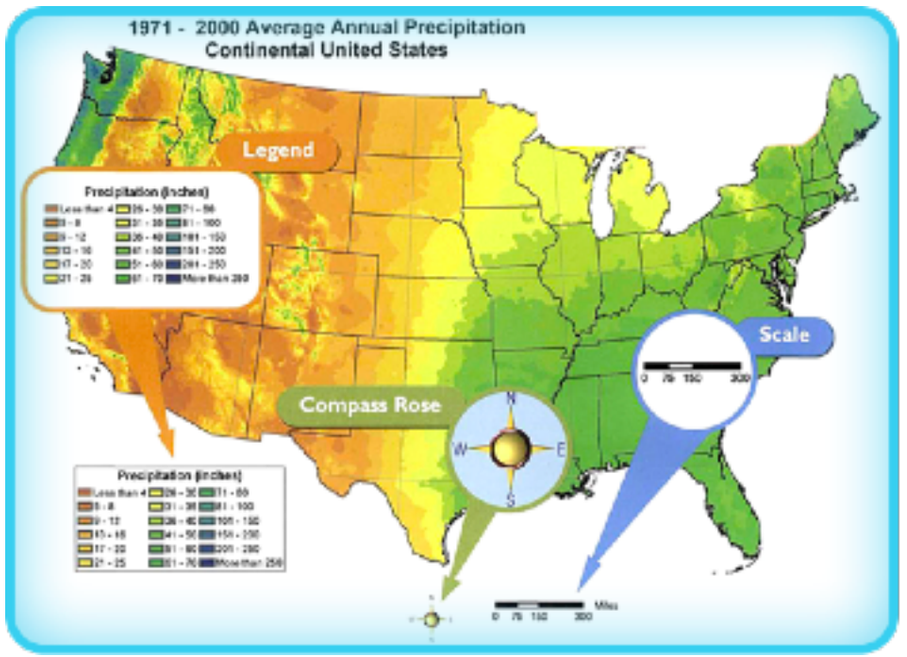
They are an alternative to a simpler approach called Fixed Sized Chunking which uses a hard coded chunk length, but has the downside of not being "shift resistant", meaning when new data is inserted into the file, all chunk boundaries to the right of the insert are changed.

Here is an example. First a file is split up into chunks of roughly even size based on content using a CDC algorithm.

New data is inserted into the middle of the file, causing everything to the right of the insert to get shifted.

File is split up again using the CDC algorithm from before. 4 out of 5 chunks match.

This new chunk is the difference between version 1 and version 2 of the file.



GIS 3: Referencing Specific Versions of Data

Type the following in the **virtual terminal** in the browser window:

```
$ dat share
```

This will create a link, that looks like `dat://...`. Your output in the **virtual terminal** will look something like this:

```
$ dat share
Created new dat
dat://5a4575c632d1a573...
```