**Principles of Big Data Management Assignment**

**Implementation of code to load data and print top 10 states:**

1. **Importing Necessary Modules**: The code begins by importing the **os** module which provides a portable way to interact with the operating system, and defining a function **list\_files()** that lists all the files in a given directory.
2. **Listing Files**: The **list\_files()** function takes a directory path as input, iterates through all the files in that directory, and appends the filenames to a list **files** if they are files (not directories).
3. **Folder Path and Files in Folder**: It defines a folder path **folder\_path** and then calls the **list\_files()** function to get a list of files in that folder.
4. **Filtering Files for Year 2021**: It creates a list **year\_2021** to store data for each month. It iterates through the files in the folder, filters out the files with names ending with "2021.csv", and organizes them into separate lists based on the month they represent.
5. **Extracting Data**: It extracts desired columns ("Province\_State" and "Case\_Fatality\_Ratio") from each file's content for each day of the month, storing the data in a nested list **final\_data** organized by month and day.
6. **Combining Monthly Data**: The next segment of the code combines the data of individual days for each month into a DataFrame for each month. This is done by Map Reduce function in pyspark.
7. **Summarizing Yearly Data**: It creates a dictionary **year\_data** where each state is a key and the corresponding value is a list of Case Fatality Ratios (CFR) for each month. It calculates the sum of CFR for each state for the whole year.
8. **Printing Results**: It iterates through the **year\_data** dictionary, prints the state name along with the total CFR for the year, and then adds a line of asterisks to separate the results for each state.

Overall, the code reads data files for COVID-19 reports for the year 2021, extracts specific columns, aggregates the data by state, and prints the top 10 highest Case Fatality Ratio for each state for the entire year.

**Output Screenshot:**

A computer screen shot of a number

Description automatically generated

**Plotting the Map:**

* The script generates a choropleth map using Plotly to visualize the COVID-19 Case Fatality Ratio (CFR) by state.
* It aggregates city-level data to the state level using a predefined mapping from cities to states (city\_to\_state dictionary).
* Plotly's Choropleth object is utilized to create the map, with each state represented by a color based on its CFR value.
* Customized layout settings are applied to the map, including title, geographic scope, and color scale.
* The resulting map provides a clear visual representation of CFR variation across different states in the USA, aiding in the understanding of the pandemic's impact.

**A map of the united states

Description automatically generated**