CS 649 Intro to Big Data
Spring Semester, 2022
Assignment 3

San Diego State University -- This page last updated on 4/12/22

Due April 28, 23:59

1. Using Spark, produce a dataset of the number of new cases of Covid-19 in each California county per week. Weeks start on Sunday and end on Saturday. Only include full weeks.

Instructions

Turn in a jupyter notebook with all the code used to answer the question. The code for problem one should be in a function that you run on AWS. To show that you ran the code on AWS, include in your notebook the AWS CLI export command for the job that you run on AWS.

Data

The data used is from USAFacts (https://usafacts.org). The data sets can be downloaded on the assignment page of the course website.

covid_confirmed_usafacts

This data set from USA Facts contains the new covid cases in each county in the country each day from the beginning of the pandemic.

URL: https://static.usafacts.org/public/data/covid-19/covid_confirmed_usafacts.csv

Column Labels

State

> 2 letter abbreviate for the State

County Name

> Name of the County

stateFIPS

> Federal ID number for the state

countyFIPS

> Federal ID number for the county

2020-01-22, 2020-01-23, etc

> The column contains the number of new cases on the specified day.

Grading

	Points
Problem 1 using Spark	15
Run on AWS	10
Only need to edit paths to test data in one location	5
No hard-coded dates in source	5

What to turn in

You need to turn in a zipped version of your python files.

Late Penalty

An assignment turned in 1-7 days late will lose 5% of the total value of the assignment per day late. The eight-day late penalty will be 40% of the assignment, the ninth-day late penalty will be 60%, and the penalty will be 90% after the ninth day late. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.