Big Data Computing

Lab - V

August 31, 2021

I. Fraud Customer Detection

Fraud risk is everywhere. One major sector affected by fraud risk is the e-commerce industry. E-commerce companies tend to collect a vast amount of data on individuals and their transactions and use modeling to distinguish between fraudulent and non-fraudulent behaviors. In this context, you are provided with a transaction dataset of an e-commerce company where each tuple of the dataset holds the information concerning following attributes:

Column	Datatype	Description
CustomerID	String	Unique ID of the user
CustomerName	String	Name of the customer
OrderDate	Date	The date when the user bought the item
ShipDate	Date	The date when the seller shipped the order
CourierName	String	Name of the courier service used for shipping
ReceivedDate	Date	The date of order receipt by the user
IsReturned	Boolean	Whether the item is returned or not by the user
ReturnedDate	Date	The date of return the item
ReasonForReturn	String	Cause of returning the item

Now, you are provided with source code(s) and instruction file that helps to output the list of customers which most likely possess the tendency of placing fraudulent orders using hadoop and map-reduce..

II. Earthquake data analysis

This data set is taken from USGS(U.S Geological Survey). The USGS provides reliable scientific information to describe and understand the Earth; minimize loss of life and property from natural disasters; manage water, biological, energy, and mineral resources; and enhance and protect our quality of life. As part of it's program, USGS monitors and reports on earthquakes, assesses earthquake impacts and hazards, and conducts targeted research on the causes and effects of earthquakes. The USGS provides real-time notifications, feeds and web services about earthquakes just after they happen. The data set contains details of all earthquakes that have happened in the last 30 days and is updated every 15 mins on the USGS website. Each tuple of the dataset holds information regarding the following attributes:

Column	Description
Src	The network that originally authored the reported location of this event.
Eqid	An identifying code assigned by - and unique from - the corresponding source for the event.
Version	Version of API that generated feed.
Datetime	Time when the feed was most recently updated.
Lat	Decimal degrees latitude. Negative values for southern latitudes.
Lon	Decimal degrees longitude. Negative values for western longitudes.
Magnitude	The magnitude for the event.
Depth	Depth of the event in kilometers.
NST	The total number of seismic stations used to determine earthquake location.
Region	Textual description of named geographic region near to the event. This may be a city name, or a Flinn-Engdahl Region name

You are provided with source code(s) utilizing **map-reduce** programs, to output the list of regions sorted alphabetically with their highest corresponding depths of earthquake events occurrences.