TEXATA WORLD CHAMPIONSHIP

WORLD'S BIG DATA SHOWDOWN

Round 2 – Big Data Case Study Information Pack

Big Data Case Study

The Question

Using any of the data listed below, as well as any other publicly available datasets you may choose to incorporate, you must produce an analysis which demonstrates the value of these datasets. There is one specific question Thomson Reuters would like answered:

Is it possible to build a set of parent-child relationships for all companies for any point in time? For example, we know that today CorpSmart is owned by Thomson Reuters, but prior to June 20 2011, CorpSmart was owned by Deloitte. Can we discover this information in the data? A sample output would be a list of Date,Purchaser,Purchasee 3-tuples. In the case mentioned above, the relevant lines would look like:

```
null,Deloitte,CorpSmart
2011-06-20,ThomsonReuters,CorpSmart
```

However, you are welcome to define your own use case based on what you understand of Thomson Reuters' business, and what is possible with the data and time available to you.

You will be required to submit a report on your analysis. These sections will appear as individual questions in the HackerRank platform:

11. **Output:** Upload the output of your analysis. This should be a file containing the results of your data analysis.
12. **Analysis Summary:** A summary of your analysis, covering the use case you chose, your results, and the significance of your results.
13. **Analysis Use Case:** A description of the use case you have chosen, explaining what business outcome you were supporting, and why that was the most valuable option for you to tackle for the round.
14. **Analysis Methodology:** A description of the methodology you used in your analysis.
15. **Analysis Outcomes:** What is the business impact of your findings? Explain the outputs of your analysis and how it would be integrated into an existing business process.
16. **Code:** Upload the code you wrote to produce the outcome.

Accessing the Data

The data for Round 2 is being stored in Amazon S3. We have prepared Amazon Web Services user credentials for all participants to access the data. The credentials to access the data are:

User name: texata-participant

Access key: AKIAJ6IGQJNSXL5RLRNA

Secret access key: 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

These credentials must be used to access the S3 buckets in which the data sets are stored. If you intend to download the datasets yourself, **you should use these credentials in your S3 download client (e.g. the Amazon AWS CLI or Cyberduck)**. If you intend to process the data using Amazon Elastic MapReduce or EC2, you must use your own AWS account credentials to create the cluster or virtual machines, and provide the above credentials when connecting to the S3 bucket.

TEXATA WORLD CHAMPIONSHIP

WORLD'S BIG DATA SHOWDOWN

Round 2 – Big Data Case Study Information Pack

Data Sets

Thomson Reuters has provided two major datasets for TEXATA 2014:

1. A sample of approximately 60,000,000 events from the Reuters News Archive. These files represent news events covered by the Reuters news network, and cover a breadth of topics, subjects and languages.

    a. The Reuters News Archive can be accessed on Amazon S3 at the bucket
    `s3://texata-2014-round-2`
    b. The bucket contains gzip-compressed UTF-8 encoded csv files.
    c. There is one file per month for every calendar month from 2004 to 2014.
    d. There are several types of events in each file, listed in the "EVENT_TYPE" field. We are most interested in "STORY-TAKE-APPEND" and "STORY-TAKE-OVERWRITE" events, but "HEADLINE" events may contain useful signal.
    e. Each file contains stories in many languages. The language is encoded in the "LANGUAGE" field.

2. A sample of approximately 140,000 earnings call transcripts from the Reuters StreetEvents database. These files are transcripts of actual phone conversations between company directors, shareholders and journalists.
    a. The Thomson Reuters StreetEvents data can be accessed on Amazon S3 from the bucket
    `s3://texata-thomsonreuters-streetevents`
    b. The bucket contains uncompressed UTF-8 encoded XML files.
    c. Each file contains a master "Event" element with several sub-elements. The most interesting sub-elements are "EventStory", which contains the transcript of the call, and "companyName", which states the name of the company which hosted the call. "startDate" is the date the call took place.

3. TEXATA has prepared a sample of 66 companies and their acquisitions giving 1662 parent-child relationships. This dataset includes the value of the acquisition where available and the jurisdiction of the parent company. This dataset could be used as a training set for a model, or a reference set for historic relationships. You can download this file from:
    https://texata-2014-round-2-acquisitions.s3-us-west-1.amazonaws.com/texata-list-acquisitions.csv