|  |
| --- |
| ***Big Data Analytics - Team Project Proposal*** |
| ***Part 1. General Project Information*** |
| **Name:** Jinglin Wang, Yicong Tao, Xi Xiao  **Project Title:** Most Prospering Business Prediction Based on Big Data Analysis  **Project Description: *(Write one paragraph to describe what this analytic will do.)***  In this project, we target at offering meaningful information and suggestions on the most prospering business in given city in the USA. By combining data from the following sources: Yelp reviews, government sales census and Twitter tweets, we aim to analyze best business choice for companies to turn their effort to in the given city in the USA. Technologies such as Hadoop, Spark and HBase will be used to store and analyze large datasets. Analytical methods such as natural language processing, N-Gram analysis, and machine learning are applied to draw insightful conclusions.  **Who is a typical user of your application:**  Startup companies, companies might want to turn to a new business.  **What insight will you derive from the data?**  Most Prospering Business in given city in the USA.  **Describe how you will prove the goodness of your analytic, in other words, how will you verify that it is correct:**  Prediction is generally difficult to verify due to its long-time window. Real-life business models can be a strong evidence in proving the analysis result. |

|  |  |  |
| --- | --- | --- |
| ***Big Data Analytics - Team Project Proposal*** | | |
| ***Part 2. Data Source Information*** | | |
| ***Name of Data Source 1:*** Twitter tweets (https://dev.twitter.com/overview/api)  ***Data Source Description:*** Social network discussion and review can be a useful information source predicting customer behaviors and popular business models, helping a customized and precise evaluation of the selected business model in certain areas. | | |
| **Data Collection Frequency**   * Are you collecting the data in realtime? Or are you collecting it periodically? * Are you collecting static data? (e.g. historic data that you load once) | **Data Size**   * *Estimate size of the data you will store, e.g. MB, GB, TB, PB* | **Data Frequency**   * If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? * If batch data, how often will you collect it? |
| ☐ Realtime (ongoing near-realtime collection)  ☑ Batch (multiple non near-realtimecollections)  ☐ Static (one time collection) | ☐ MB  ☐ 1-10 GB  ☑ 10-100 GB  ☐ 100-300 GB  ☐ 300-500 GB  ☐ > 500 GB | **If realtime data:**   * How often will you collect data?   ☐ Every second, or every few seconds  ☐ Every minute, or every few minutes   * What is the size of data you will collect at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **If *not* realtime data:**   * Will you collect a batch of data periodically pr just once (static)?   ☐ Just once  ☐ Every hour, or every few hours  ☑ Every day, or every few days, or every week, or every month   * How much data that will be collected at each interval?   About 1 ~ 10 GB in each interval |

|  |  |  |
| --- | --- | --- |
| ***Big Data Analytics - Team Project Proposal*** | | |
| ***Part 2. Data Source Information*** | | |
| ***Name of Data Source 2:*** Yelp data set (https://www.yelp.com/dataset\_challenge)  ***Data Source Description:*** The data set includes 2.7M reviews and 649K tips by 687K users for 86K businesses. | | |
| **Data Collection Frequency**   * Are you collecting the data in realtime? Or are you collecting it periodically? * Are you collecting static data? (e.g. historic data that you load once) | **Data Size**   * *Estimate size of the data you will store, e.g. MB, GB, TB, PB* | **Data Frequency**   * If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? * If batch data, how often will you collect it? |
| ☐ Realtime (ongoing near-realtime collection)  ☐ Batch (multiple non near-realtimecollections)  ☑ Static (one time collection) | ☐ MB  ☑ 1-10 GB  ☐ 10-100 GB  ☐ 100-300 GB  ☐ 300-500 GB  ☐ > 500 GB | **If realtime data:**   * How often will you collect data?   ☐ Every second, or every few seconds  ☐ Every minute, or every few minutes   * What is the size of data you will collect at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **If *not* realtime data:**   * Will you collect a batch of data periodically pr just once (static)?   ☑ Just once  ☐ Every hour, or every few hours  ☐ Every day, or every few days, or every week, or every month   * How much data that will be collected at each interval?   About 3 GB |

|  |  |  |
| --- | --- | --- |
| ***Big Data Analytics - Team Project Proposal*** | | |
| ***Part 2. Data Source Information*** | | |
| ***Name of Data Source 3:*** Geographic Area Series: Economy-Wide Key Statistics: 2012  (http://www.census.gov/)  ***Data Source Description:*** The data are shown for employer establishments at the US, State, Combined Statistical Area, Metropolitan and Micropolitan Statistical Area, Metropolitan Division, Consolidated City, County (and equivalent), and Economic Place (and equivalent; incorporated and unincorporated) levels for the U.S. and the Island Areas. Data for nonemployer establishments are shown for the U.S. for all levels except Economic Places and only for Puerto Rico for the Island Areas. | | |
| **Data Collection Frequency**   * Are you collecting the data in realtime? Or are you collecting it periodically? * Are you collecting static data? (e.g. historic data that you load once) | **Data Size**   * *Estimate size of the data you will store, e.g. MB, GB, TB, PB* | **Data Frequency**   * If realtime data, what is the frequency and volume of data (how often and how much data will you collect at a time)? * If batch data, how often will you collect it? |
| ☐ Realtime (ongoing near-realtime collection)  ☐ Batch (multiple non near-realtimecollections)  ☑ Static (one time collection) | ☐ MB  ☑ 1-10 GB  ☐ 10-100 GB  ☐ 100-300 GB  ☐ 300-500 GB  ☐ > 500 GB | **If realtime data:**   * How often will you collect data?   ☐ Every second, or every few seconds  ☐ Every minute, or every few minutes   * What is the size of data you will collect at each interval?   \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **If *not* realtime data:**   * Will you collect a batch of data periodically pr just once (static)?   ☑ Just once  ☐ Every hour, or every few hours  ☐ Every day, or every few days, or every week, or every month   * How much data that will be collected at each interval?   2.27GB in total. |