|  |
| --- |
| ***Big Data Application Development - Initial Project Proposal*** |
| ***Part 1. General Information*** |
| **Submitted by:**  CHEN-YUAN HO  **Project Title:**  I Wanna have Affordable Housing! Factors of Housing Trend.  **Project Description: *(Write one paragraph to describe what this application will do)***  To begin with, I will collect the following four historical data. The first is dataset of housing cost, including rental and sales price, in New York City. The second is dataset of CPI in New  York City. The third is dataset of property tax rate in New York City. The fourth is dataset of employment rate. Then, I move to the second step. I will use Spark in Hadoop environment  to do ETL and data profiling. Then, I will run statistical tool, such as Chi-Square, Correlation and Coefficient in Spark to know the relationship among factors in housing cost. In the end, I  will try to build a linear regression model for housing cost based on factors of CPI, property tax and employment rate.  **Who is a typical user of your application:**   1. Government 2. Tenant 3. Landlord 4. Leasing Offices 5. Real Estate Agent / Company   **What insight will you derive from the data?**   1. The relationship between CPI and housing cost. 2. The relationship between property tax rate and housing cost. 3. The relationship between employment rate and housing cost.   **What actuation decisions can be made based on the actionable insight?**  Implement policies related to CPI, taxation and employment to reducing the housing cost. |

|  |
| --- |
| ***Big Data Application Development - Initial Project Proposal*** *(continued)* |
| ***Part 2. General Data Source Information*** |
| **Name of Data Source 1:** |
| **Data Source Description** *(Provide a short description of the data source.)* |
| **Data Size** (Estimate size, e.g. MB, GB, TB?) |
| **Data Collection Frequency**  Is the data source a static, periodic, or realtime (i.e., near realtime) source?  Static  If realtime data, what is the frequency with which you will collect the data and what is the volume of data collected at each interval?  N/A  If not realtime data, will you collect a batch of data periodically pr just once (static)?  Just once, because it is historical data.  If the data will be collected periodically, how often will you collect it and what is the volume of data that will be collected at each interval?  N/A |

|  |
| --- |
| ***Big Data Application Development - Initial Project Proposal*** *(continued)* |
| ***Part 2. General Data Source Information*** (continued) |
| **Name of Data Source 2:** |
| **Data Source Description** *(Provide a short description of the data source.)* |
| **Data Size** (Estimate size, e.g. MB, GB, TB?) |
| **Data Collection Frequency**  Is the data source a static, periodic, or realtime (i.e., near realtime) source?  If realtime data, what is the frequency with which you will collect the data and what is the volume of data collected at each interval?  If not realtime data, will you collect a batch of data periodically pr just once (static)?  If the data will be collected periodically, how often will you collect it and what is the volume of data that will be collected at each interval? |