Big Data Analytics - Team Project Proposal

Part 1. General Project Information

Names of team members: Ankit Bhagat, Sanketh Purwar

Project Title: Visualizing how Airbnb is accommodating people around the world.

Project Description: ***(Write one paragraph to describe what this analytic will do.)***

This project is divided into three steps: Collecting data, Pre-processing data and Visualizing the data. We would be collecting data in batches from http://insideairbnb.com/get-the-data.html. The next step would be pre-processing the data we obtain from the website in CSV format, store it into a database. The end product/output would be a visualization platform that can be used to have a better understanding of the data in a more abstract manner. The end product/output will help with predicting the availability of a particular room next year or next month. This can act as a business intelligence (BI) tool and a tool for Airbnb customers to actually analyze the various areas of the city.

Who is a typical user of your application:

Analysts at Airbnb and visitors trying to select a place to live in the city (analyzing the visual data, instead of going through the reviews)

What insight will you derive from the data?

Detailed listing, prices, geography, neighborhood, availability, reviews of different Airbnb locations.

Describe how you will prove the goodness of your analytic, in other words, how will you verify that it is correct:

The analytical part should be self explanatory. We would try to verify its correctness by comparing it against the data present online to the public/ customers.

Big Data Analytics - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 1: Airbnb NYC Data

Data Source Description: Data source consists of various comma separated value files. The files contain data about the reviews, geography, calendar, detailed listing, etc. for the Airbnb locations in New York City, NY.

Data Collection Frequency	Data Size	Data Frequency
• We are working on static data present online: http://insideairbnb.com/get-the-data.html	• *Estimate size of the data you will store, e.g. MB, GB, TB, PB*	• 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 or 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? Airbnb releases its data on 2nd of every month (and archives the data). We would try to collect data every month. We would be collecting around 1 GB of data at every interval. _____

Big Data Analytics - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 2: Airbnb SF Data

Data Source Description: Data source consists of various comma separated value files. The files contain data about the reviews, geography, calendar, detailed listing, etc. for the Airbnb locations in San Francisco, CA.

Data Collection Frequency	Data Size	Data Frequency
• We are working on static data present online: http://insideairbnb.com/get-the-data.html	• *Estimate size of the data you will store, e.g. MB, GB, TB, PB*	• 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 or 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? Airbnb releases its data on 2nd of every month (and archives the data). We would try to collect data every month. We would be collecting around 1 GB of data at every interval. _____

Big Data Analytics - Team Project Proposal

Part 2. Data Source Information

Name of Data Source 3: Airbnb Paris Data

Data Source Description: Data source consists of various comma separated value files. The files contain data about the reviews, geography, calendar, detailed listing, etc. for the Airbnb locations in Paris, France.

Data Collection Frequency • We are working on static data present online: http://insideairbnb.com/get-the-data.html	**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 or 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? Airbnb releases its data every few months for Paris. We would try to collect data whenever they do. We would be collecting around 1 GB of data at every interval. _____