# Step 1: Review Hadoop Project Video and Answer Questions

a. Selected Project:

Juvo – Assisting local community leaders achieve their social and economic development goals

b. Answered Questions:

1. What were the goals of the selected project?

For the Juvo project, the creator’s main goal was to provide the target users a way to use realistic goals and metrics during the process of setting social and development plans. In addition, the creator wanted to determine if events such as homicide, suicide, and population decline harm a community’s quality of life and/or suppress economic development. By creating the Juvo app, the target users (policy makers, public administrators, and economic development officials) would have a method of gaining insights about their community’s characteristics. The creator also wanted to set a baseline score for communities relative to others around the world, providing a yardstick to measure progress.

1. Describe the data sources used in the project and evaluate whether these data sources meet the criteria to be described as "big data."

Three data sources were used in this project, each from a department within the United States (US) federal government. This government data is publicly available and found on the data.gov website. First, from the US Treasury department, migration data was used which is tracked through addresses from individual tax returns that are filed from one year to the next. The second, from the US Health and Human Services department (HHS), mortality data was used which is based on all death certificates filed in the US. Lastly, from the US Commerce department, local community characteristic data was used which is tracked through the American Community Survey. These data are maintained by IRS, National Center for Health Statistics, and the Census Bureau, respectively.

When evaluating these data sources against the five “big data” criteria (volume, velocity, variety, value, veracity), it could be argued that the data is not actually “big data”. However, for those that evaluate big data on volume alone then all three datasets would be classified as “big data” due to the size of dataset and amount of data that is being reported. As mentioned, the amount of data from each data source was voluminous; however, one could argue that the data may be more structured than unstructured since it was coming from federal and/or state forms and even the survey since it would be more question and choice format versus free text fields. Although death certificates are not federal forms (i.e., held by National Archives and Records Administration) and can vary by state, the forms still have structured data fields versus semi-structured or unstructured fields such as a free-text field. Additionally, the veracity factor for the datasets could fall under big data since you may have people adding street or state names/abbreviations differently, some may use all caps and others may use initial cap while others could even use lowercase or some mixture of these. That is just one example but there could be many more within the datasets. Within the three data sources there is much data that can be reduced to bring value to an end user. After considering the above items, I would state that these data sources do meet the criteria described as “big data”.

1. What analytical methods were used in this project, and why was Hadoop needed to perform this analysis?

In my opinion, five different analytical methods were used in this project. The first being descriptive data analysis which outlined items such as constant economic decline and target users repeating same mistakes regarding economic development as well as the virtuous cycle occurring when economic/social development are focused, inclusive, and evolving. After determining what happened, the creator moved onto the exploratory data analysis phase. This phase explored the data to find relationships/connections and, from the IRS migration data, discovered two outlier communities with the largest migration outflows and then dove further into the detail and reviewed the number of migrations for those selected outliers. Next phase focused on the diagnostic data analysis, where the creator tried to determine why New Orleans 2005 and Los Angeles (LA) 2004 to 2007 were the two outliers as well as why LA had three times more people who migrated than New Orleans. This analysis phase was also trying to determine the death rates by geographic classification. Following the diagnostic analysis phase is the predictive data analysis phase where the focus changes to what will happen. Within this phase, the creator was trying to understand how combining ranking data with American Community Survey data would affect the community characteristics as well as determining the floor (bottom 10%) and ceiling (top 10%) metrics of the Workforce, Social Economic, and Behavioral groups by their specified characteristics. This brings us to prescriptive data analysis which is the last data analysis phase for this project and is where the ‘how will it happen’ question is answered. Here the creator shows how their app, Juvo, will aid and assist the target users when setting their development plans. It provides the ability to explore the data by category (Workforce, Social Economic, Behavioral) and sort data by the detailed group characteristics, giving users a way to not only set realistic and achievable goals but also metrics to measure those goals.

Within this project, the creator had a need to use Hadoop for determining the effect of missing data on analysis, providing added insights into the data with key missing aspects included. In addition, the raw mortality data alone was greater than one gigabyte and they needed the capacity to count the total number of homicides and suicides from this large dataset.

1. Describe how the results of the analysis were presented (i.e., via maps, charts, tables, etc.), and evaluate how effective you believe this presentation was.

Within this project, analysis results were presented using charts such as R studio’s ggplot2 graphs for understanding overall migration over time and LA migration outflows, box plots by county from combined migrations and mortality data, bar charts featuring the top and bottom ten percent communities and data tables providing a deeper dive into the characteristics of the three different category groups (Workforce, Social Economic, Behavioral).

Overall, I believe this project met its main goal of providing target users a way to use realistic goals and metrics when setting social and development plans. However, there are items in the presentation that leave me with questions and may have been summarized too much. For example, why does the Behavioral group only include males, how does this app show suppression of economic development from the specified factors, is this providing the baseline score relative to others around the world?

1. What are some ways in which this project could potentially be extended or improved?

This project could be improved by updating the behavioral group characteristic definitions to avoid leaving out single moms with daughters and even single dads with daughters, confirming a broader group was accounted for or adding more characteristic groups. As for extending the project, adding elements to show how these items affect economic development or providing the baseline score relative to other communities around the world would provide additional benefit to the target users.

1. What other applications can you envision using the analytic methods and/or data sources involved in this project?

Descriptive analytics can be used to collate survey results, report general trends, or summarize past events such as sales and operations data. Predictive analytics can be used to forecast trends across industries/businesses as well as the identification of possible security breaches that requires further investigation by IT. Prescriptive analytics can be used to assess risk in regard to pricing of insurance premiums, identify the best testing patient groups for clinical trials, and track fluctuating prices of oil/gas or stocks.

c. Hadoop Learnings

1. Simplilearn Tutorials
   1. Lesson 1: What is Hadoop? (<https://www.simplilearn.com/tutorials/hadoop-tutorial/what-is-hadoop?source=sl_frs_nav_playlist_video_clicked>)

Great video tutorial! Loved the farm analogy and how the components were laid out and discussed.

Graphical user interface, application

Description automatically generated

* 1. Lesson 7: HDFS Tutorial (<https://www.simplilearn.com/tutorials/hadoop-tutorial/hdfs?source=sl_frs_nav_playlist_video_clicked>)

Diagram

Description automatically generated

1. Hadoop Tutorial: 5 Real-World Examples of Hadoop (<https://www.bmc.com/blogs/hadoop-introduction/>)
   1. Financial services companies use analytics to assess risk, build investment models, and create trading algorithms; Hadoop has been used to help build and run those applications.
   2. Retailers use it to help analyze structured and unstructured data to better understand and serve their customers.
   3. In the asset-intensive energy industry Hadoop-powered analytics are used for predictive maintenance, with input from Internet of Things (IoT) devices feeding data into big data programs.
   4. Telecommunications companies can adapt all the aforementioned use cases. For example, they can use Hadoop-powered analytics to execute predictive maintenance on their infrastructure. Big data analytics can also plan efficient network paths and recommend optimal locations for new cell towers or other network expansion. To support customer-facing operations telcos can analyze customer behavior and billing statements to inform new service offerings.
   5. There are numerous public sector programs, ranging from anticipating and preventing disease outbreaks to crunching numbers to catch tax cheats.
2. Blog with 12 commands and Practice Examples (<https://data-flair.training/blogs/hdfs-hadoop-commands/>)

Loved learning more about Hadoop Distributed File System!

Diagram

Description automatically generated with medium confidence

# Step 2: Investigate an Organization that Uses Hadoop

Selected Organization

Adobe

Brief Description of Adobe

Adobe connects content and data as well as introduces new technologies that democratizes creativity, shapes the next generation of storytelling, and inspires new categories of business. This is accomplished through many different applications falling under the umbrella of Adobe Creative Cloud. Creative Cloud is a collection of 20+ desktop and mobile applications and services for photography, design, video, web, and more. Businesses and students count on Adobe to help them meet challenges of digital transformation and accelerate document productivity.

Summarization of How Adobe Uses Hadoop

Adobe has been using Apache Hadoop and Apache HBase since October 2008. They are used in several areas from social services to structured data storage and processing for internal use. Adobe has approximately 30 nodes running the Hadoop Distributed File System (HDFS), Hadoop, and HBase in clusters ranging from 5 to 14 nodes on both the production and development environments. Deployments are typically planned on an 80‑node cluster. Adobe constantly write data to Apache HBase and run MapReduce jobs, processing then storing back to Apache HBase or external systems.