Following the Tech Industry

Silicon on the Move

Joshua Tyndale
Computer Science Dept
Colorado State University
Fort Collins, USA
Jtyndale@rams.colostate.com

Stephen Harayda
Computer Science Dept
Colorado State University
Fort Collins, USA
sharayda@rams.colostate.edu

Dziugas Butkus
Computer Science Dept
Colorado State University
Fort Collins, USA
dziugas@rams.colostate.edu

Keywords—Apache, Spark, MapReduce, PySpark, Data

I. Introduction

A large part of graduating college is knowing where to apply to your first job. Many factors can go into this decision, including but not limited to, job availability, economic growth of the city, which companies and what job types are specific to that region, and other various personal preferences. Using pySpark, an in depth analysis was done to both job, and housing data sets, from indeed, and craigslist, respectively, to find the best places for computer science students, or anyone in the tech industry. Finding, and going to an area that is already established, or has growth can help a new grad find a plethora of jobs, with a variety of choices in professional development. This type of research can be useful to anyone in search of finding employment, and this type of data analysis is done all the time for journalists writing articles about the next big tech hub.

Using a master node and set of worker nodes, Apache Spark, integrated with python (pySpark), was used to break these data sets down into workable Resilient Distributed Datasets (RDD's). Several calculations were then done on the data, such as filtering for tech jobs, counts of job occurrences per city/state, calculating the price per square foot of regions, and more. After these programs were run, the highest occurring attributes were printed out for further analysis. Using these outputs, it is possible to determine, with some error, what parts of the United States are ideal for job seekers in the tech industry.

Finding economic and technological hubs that are booming, or up and coming, can be difficult to pinpoint just by surfing the web. Most published works are biased, trying to push an agenda, or want to create a certain narrative. Many places market themselves as 'silicon something or other' to get attention from the tech world. While the analysis done here does not encompass everything that is happening around the US and with the tech industry due to problems discussed in the next section, it does offer some interesting, unbiased information that can be helpful to anyone job hunting.

While it may not be obvious, this type of work is seen everywhere. In today's internet climate, data is constantly referenced in articles, speeches, and is the determining factor in many of the policies that are put into place by government officials. Correctly analyzing data to show what is actually happening is much more difficult than it may seem.

II. Problem Characterization

There are a few problems to consider when looking to analyze the job market and knowing where to apply. These problems include access to employer information, and job data, as well as data which is timely so that it can be useful. Today more than ever companies are becoming more possessive of their data which makes it harder for people to access that data and use it for analysis. Also, because of the large amounts of data it is important to narrow the focus of the analysis so that it is useful for the person who is looking for jobs. If the analysis is too broad the amount of data can become overwhelming and make the issue of looking for jobs more difficult. Another problem when analyzing jobs and housing markets is that the data available is only a snapshot of the time it was accessed, this can cause problems because while it does offer a good picture of the job market at that time, it fails to capture where the job market is headed next. A good example of this is the current trend of people in tech moving to Miami for jobs as discussed in a podcast interview by the Economist with the Mayor of Miami, Francis Xavier Suare [1]. This trend is known by looking at news articles or tweets on twitter but the job market data has not yet captured these trends.

Another issue when analyzing the job market is finding places to live that not only offer an abundance of jobs in the given field but also offer places to live for a reasonable price. Combining job market datasets with housing datasets can be tricky because it is important that they overlap. For example, if most of the jobs available are in big cities then it is important to analyze apartment prices for these places in question. Not only do these two datasets need to overlap in terms of location they also need to correlate in terms of time accessed so that the job market correlates with the given prices for apartments.

The final issue when analyzing where to look for jobs is the person's own personal preferences. If the analysis shows that most jobs are available in major cities but the person is looking to live in a small town then the analysis might miss out on creating a picture which is useful for that given person. Also the person might not care about where the jobs are located or the prices of apartments but rather other ratings such as night life, transportation or factors which relate to quality of life.

III. Dominant Approaches to the Problem

There have been multiple approaches to the problem of identifying the best places to live when looking for jobs in the tech industry. These approaches include ranking cities based on cost of living and expected income, finding cities based on total job postings, and using surveys on people in the tech industry to find out information as to how much and where they live. Each of these approaches offer their own set of advantages and disadvantages when considering where to live and apply to work.

The first approach of looking at jobs based on cost of living and expected income was done by Bloomberg. Their approach was to find out salaries based on data from Indeed and use the U.S. Bureau of Economic Analysis to find the cost of living for a given region [2]. The advantages of this approach is that they are able to identify where future employees can look to get the most out of their money. The disadvantages of this approach is that it lacks an outlook on where the available jobs tend to be. A city may have a higher adjusted salary but there may not be very many jobs available in that region.

The next approach was to find the cities with the most total job postings in the tech industry. This analysis was done by Indeed and focuses the most on where job availability is currently located [3]. The advantage to this approach is that it shows where to look for jobs as well as the skills these employers are most interested in. The disadvantage to this analysis is the opposite of the Bloomberg analysis in that it fails to take into account how much a person can expect to pay when living in a given city.

The final approach was done by PCMag and it used a survey performed by Dice where over 10,000 employees across the country were asked to identify how much they make and then adjusted it for cost of living [4]. It also asked employees to identify how satisfied they were with their given salary. This approach has the advantage of giving a good picture of where employees can expect to make the most money and be satisfied where they work. By doing this it offers a better insight into quality of life in a given location. The issue with this approach however is that it fails to account for where employers are currently hiring.

IV. Your methodology

Spark is compatible with Python, Scala, SQL, and R, although Python was used in this experiment due to previous work and experience with analyzing, displaying and cleaning the data using Python. Using Jupyter Notebook was the first course of action for the whole of the assignment, including the report, which would allow the assignment to be replicated in real time while reading about it. However, issues were encountered with excessive output and the time spent on setting up the cluster and environment made Jupyter Notebook an illogical choice. Due to these issues, the terminal was used for running the PySpark code.

After setting up PySpark and finding the most suitable datasets for housing and job listing markets, the data needed to be cleaned. Most of the data contained irrelevant information, like description, url, images, etc. Some of the datasets had more than 100 columns, so the schema of each DataFrame was printed and only the columns that were needed for this project were left.

Job listing dataset came with several different .json files. Working with them separately would require too much redundant code. Before joining every file together, we filtered out the information with only information technology jobs such as Software Engineer, Dev Ops, etc., and left out all other irrelevant jobs. To ensure this was done correctly, many different, carefully chosen keywords were used to filter the data.

After filtering, cleaning the data and dropping columns, there was still data that was either in the wrong place, for example, some of the descriptions were mixed up with city names, or users created real listings using fake or placeholder values that were too unrealistic. Those values were big enough to drastically change our results, thus had to be removed.

Once the data was clean, and filtered it was ready to be used for the analysis. First, the queries and questions to be performed on the data needed to be created conceptually, so that it would have meaningful information, and it would fit both datasets. Since one of the dataset is a summary of the housing market and the other is about tech job listings, paying significant attention to this task was detrimental to the success of this project. The queries that we came up with were:

Getting a count of all job listings per state and city in the US, and ranking the locations in descending order: This was done to discover what states and cities had the most economic activity and job availability. Just observing the quantity of postings alone can serve a valuable insight into the economy of a city or region.

Getting a count of all job listings per company, and ranking the companies in descending order: A list of jobs per company is useful to those who would like to see what the largest companies are, either to avoid, or to seek out.

Gathering a list of companies that were hiring the most in certain areas; Number of rental listings per region: This allows someone to evaluate what companies are in the area that they are currently in, or what companies are in the location that they want to move to.

Average monthly rent price by region: Obviously, the cost of living is a big determining factor of people's decisions to move to certain areas, and overall rent price is a large part of that. The rent price was listed in descending order, to look at lower cost of living relative to other areas that were booming. A list in ascending order would have produced a list of locations in the middle of nowhere, which certainly would not have been tech hubs, let alone viable places to get a job at all.

Average price per square foot by region: An analysis was done on price per square foot for the reason that the previous, overall rent-based analysis, does not take into account the size of the rental. A house available for $2,000 a month would be a much different experience than an apartment that is available for $2,000 a month. This also takes into consideration rooms per rental, as with more rooms means more space.

And lastly, the most cat-friendly, and most dog-friendly places to live: Moving can be a lonely process, and a companion can be helpful. Pets being allowed into rental spaces can be a big determining factor of a desirable location. To check for this, the cities with the most cat and dog approvals were found.

Using all of these queries, a list of best regions could be formed. No one single query alone can fully encapsulate the best regions in the United States to look for employment, but by combining all of these outputs together, it is possible to form a reasonable list of economic hubs where someone in the tech industry could thrive.

With all of the data analysis planned, the next step was to find a good source of information on how to work with and analyse pySpark DataFrames. One of the most useful resources found was 'sparkbyexamples.com/pyspark/' [5]. This resource, combined with other random hints from forums, made it much easier to get started with writing programs in pySpark. Luckily, dataFrames come with a

plethora of functions such as .filter(), .groupBy(), .orderBy(), .count() and more.

Once all of the outputs were generated and gathered, they were combined to a single location for ease of access and analysis. Jupyter Notebook was then used to graph results for visualization of the conclusions reached. This was done to create ease of reading by a wide audience, and visual break from words and code.

The final part of the evaluation was done by hand. Each member of the team evaluated all of the graphs, and chose which cities were best for job hunters in the technical industry. The first graph took into consideration were the cities with the most amount of tech jobs posted. After that, cost of living, and most readily available housing were looked at to narrow the list down. For example, New York had the most amount of tech jobs, but after looking at available rental properties, and cost of living, it was taken off the list of best cities to move to.

V. Experimental benchmarks

At first, before any sources were cross-referenced, the outputs given by the experiments matched up with the preconceived notions held, of what the biggest hubs were, and what the most expensive places to live were.

The quality of the data was then verified by referencing other sources online. According to the Indeed career guide, some of the best cities for software engineers include Austin, Dallas, Atlanta, New York, cities in California, and others [6]. This list fits very closely with the list of cities that was put together by the output produced in this experiment. While it is not always exact, there are many different variables at play that determine these lists, which are discussed above. The data created in this experiment did produce a few outliers that were not relevant, such as remote jobs, which have no physical location, and were not included in our analysis.

Additionally, the housing data output was cross referenced with references online, one being move.org [7]. The list that they have put together of most expensive cities per square foot matches the list created in this experiment very closely. While it is not ideal to live in the most expensive cities in the US, the job data was referenced first and foremost, to determine the best locations. After referencing the job data results, the housing data was then evaluated for cost of living as a secondary factor.

The only other benchmarks used to ensure that the evaluation was correct, was making sure that counts and attributes made sense. For example, when unions were performed on multiple data files, it was verified that the totals for each attribute added up accordingly, and when the highest numbered attributes were very low, the logic behind the source code was evaluated for correctness until the expected, and desired output was given.

VI. Insights gleaned

Before beginning this project, the only prior knowledge known about Apache Spark was the general concept. Without any prior experience, getting started was the most difficult part. Setting up the environment and getting clusters running were one of the most time consuming parts. Although, once the environment was set up and the cluster was up and running, the speed and efficiency of the program was surprising. Processing gigabytes of data took maybe 20-30 seconds at the most.

At first, an attempt was made at using Scala to run the programs, but the learning curve was quickly found to be too much. The lack of experience with the language impacted performance greatly. After looking into it, python seemed to be the best choice due to prior knowledge and experience with the language. After further research, PySpark seemed to be the best choice, so changes were made to the environment accordingly, more specifically the PATH variables in the .bashrc file for Python and PySpark. The execution command was then adjusted without any issues.

During this project, it became very clear how important the state of the data is. It needs to be clean already, or be cleaned manually before any useful analysis can be done on it. Initially, the plan was to use API's from Zillow, Indeed, and Glassdoor, to create a dataset from scratch, but due to knowledge and time limitations, a pre-made dataset had to be used. It also seems that many API's are becoming more private, and not enough resources were available for creating a dataset that was large enough. A housing dataset and a job dataset proved to be useful, but this was found after hours of attempting to use other datasets that were unusable due to not being compatible with our data structures, or had to be bought.

Before any analysis was done on the datasets, it was expected that the outputs would be similar to what information was spread around in the mainstream media as 'tech hubs' such as Texas, California, and Miami. After analyzing the data, both housing and job listing results were stereotypical. Even though housing data was a snapshot in the middle of the Covid-19 pandemic, it seemed stereotypical, to what was expected pre-pandemic and corresponded with the job listing data which was snapshotted January 2020, before the pandemic. New York, Seattle, Washington, Chicago, Austin and San Francisco were at the top of the list. Austin was very popular in job listings, but it does not appear in any of the top 20 lists of the housing data.

VII. Transformation of the problem space in the future

As data grows exponentially, new methods of performing data analytics will be developed. Data is a large driving force of our day to day world, inspiring articles, decisions, and technologies. The overall approach of MapReduce, using many worker computers to perform one computationally time demanding task, will be used for quite some time due to its place in tech companies, such as google, and due to its ease of use, and efficiency.

A trend that seems to be catching on is privacy of data. It is becoming more and more difficult to find readily available, usable data, free of charge. For example, access to APIs such as Zillow have been limited to users, and many of the free datasets found are not as encompassing for meaningful analysis. This is unfortunate for researchers and scientists, because readily accessible data can make a world of a difference. Data availability, going forward, will help scientists, which in turn, will make everyone's lives better [8]. The better the research, the better the products, and services [8].

A way to overcome the trend towards data becoming more private is to use public forums as a source of data. With the development of sentiment analysis with machine learning, researchers have been able to use public forums such as twitter and keywords to analyze current trends in problem spaces such as stock market trends [9]. These same algorithms could be used in areas like job

availability and job satisfaction. The advantage to this development is that it will better capture where the trends are heading in real time, based on up-to-date tweets.

VIII. CONCLUSIONS

There are many ways to analyze the data, and based on the approach, the results may not be indicative of the reality. A lot of those approaches may not be correct and the data may be manipulated to get the result that the subject needs. It was found that most datasets need different time periods to be able to find the change and trends in data. Without them, data is very one dimensional and is limited to conclusions at that specific point in time.

We used two different datasets, one was collected before the pandemic, and the other one was in the middle of the pandemic. While the results were not surprising and did not significantly differ, it is important to have all datasets with the same time periods to make the most accurate comparisons.

The results presented may be skewed due to recruiting pushes. For example, Tesla has very few job postings, because everyone seeks them out, and applies on their website via email, while the military has recruiters who push postings out as much as possible. During the time of our analysis the top 5 companies hiring were Amazon, JP Morgan, Navy, Compass Group, Air Force.

Based on our analysis we found that the top 5 cities students should look for jobs are Washington D.C., Seattle, Chicago, Austin, and Atlanta. These cities offered the most jobs for students in tech while also being the most reasonable in terms of prices for apartments. Cities such as New York failed to make our top list due to high rent prices. Using this analysis students can be more informed about job availability and where to look next for jobs.

Cities with the most amount of jobs accounting for rent prices:

Washington DC, Seattle, Chicago, Austin, Atlanta

Companies hiring the most:

Amazon, JP Morgan, Navy, Compass Group, Air Force, Microsoft, Army

REFERENCES

[1] A. McElvoy, “How do you reinvent a city?,” The Economist. [Online]. Available: http://www.economist.com/podcasts/2021/04/15/how-do-you-reinvent-a-city. [Accessed: 28-Apr-2021].

[2] R. Florida, Bloomberg.com, 05-Sep-2019. [Online]. Available: https://www.bloomberg.com/news/articles/2019-09-05/ranking-cities-by-salaries-and-cost-of-living. [Accessed: 29-Apr-2021].

[3] J. K. Murray, “The Top 10 Cities and Metropolitan Areas for Tech Jobs in 2020,” Indeed Career Guide, 29-Mar-2021. [Online]. Available: http://www.indeed.com/career-advice/finding-a-job/top-cities-for-tech-jobs-2020. [Accessed: 28-Apr-2021].

[4] C. Steele, “20 High-Tech Cities You'll Want to Call Home,” PCMAG, 20-May-2019. [Online]. Available: https://www.pcmag.com/news/20-high-tech-cities-youll-want-to-call-home. [Accessed: 30-Apr-2021].

[5] *Nnk, “PySpark - SparkByExamples,” Spark by {Examples}. [Online]. Available: https://sparkbyexamples.com/category/pyspark/. [Accessed: 28-Apr-2021].*

[6] *Indeed Editorial Team, “10 Best Cities for Software Engineers,” Indeed Career Guide. [Online]. Available: https://www.indeed.com/career-advice/finding-a-job/best-city-for-software-engineers. [Accessed: 27-Apr-2021].*

[7] *T. Wheelwright, “America's Most Expensive Cities per Square Foot,” Move.org, 16-Apr-2021. [Online]. Available: http://www.move.org/americas-most-expensive-cities-per-square-foot/. [Accessed: 28-Apr-2021].*

[8] “Data sharing and the future of science,” Nature News, 19-Jul-2018. [Online]. Available: http://www.nature.com/articles/s41467-018-05227-z. [Accessed: 28-Apr-2021].

[9] Y. Takahashi, “LSTM vs bert - a step-by-step guide for tweet sentiment analysis,” 09-Nov-2020. [Online]. Available: https://towardsdatascience.com/lstm-vs-bert-a-step-by-step-guide-for-tweet-sentiment-analysis-ced697948c47. [Accessed: 29-Apr-2021].