Journal of Systemics, Cybernetics and Informatics (2022) 20(2), 40-44
https://doi.org/10.54808/JSCI.20.02.40

A Study in Performing Big Data Analytics with Limited Resources

Steven MATHIS and **John COFFEY**
Department of Computer Science
The University of West Florida
Pensacola, FL, 32514

ABSTRACT[1]

Data analytics are usually the purview of large organizations with significant human and computational resources. The purpose of the article is to identify how organizations with limited resources, including limited computing power and bandwidth, can carry out meaningful data analytics at a granularity that fits their specific needs. This paper presents a case study in creating and maintaining a COVID dashboard for a 5-county area in the Panhandle of Florida, using only free software and publication platforms. Results of this study are informative for any local or regional entity needing focused data analytics.

Keywords: Data Analytics, COVID-19, automated data acquisition, automated summarization, automated update

1. INTRODUCTION

High quality data analytics are critical today to help individuals and organizations make sense of the enormous amount of data that is immediately available everywhere. The costs of data summarization and the granularity at which data are summarized are important ongoing issues. Large-scale data analytics are expensive and may not provide the proper focus or timeliness required for specific data consumers. This study came about as a retrospective from a year of maintaining a COVID dashboard for a 5-county area in the Panhandle of Florida.

The remainder of this paper will cover a review of the literature involving related studies, a discussion of motivations for the work, details on the specific methods used for data retrieval, processing, presentation, a summary of the results of the work accomplished, decisions made, and challenges faced along the way.

2. RELATED LITERATURE

Much of the literature on data analytics has a focus on large scale, enterprise-level operations. For instance Yamada and Peran [1] describe a reusable framework for data analytics governance in which they set up checkpoints in the process to foster communication and coordination between managers and analytics practitioners.

Grady et al [2] claim that data analytics process models typically follow a lifecycle resembling the waterfall model for software development. They advocate for a more agile approach to try to optimize the value of data and the time and resources required to generate actionable results.

Londhe and Rao [3] describe various platforms for data analytics including an increasing trend towards hybrid approaches that integrate multiple platforms. The authors state that choosing the best hardware/software combination for the job is an increasingly complex endeavor.

Dealing with missing data is a key issue in efforts to ensure data veracity. Ehrlinger et al [4] describe and evaluate methods to impute values for missing data. They identify several missing data patterns and conclude that missing data patterns are domain-specific.

Data analytics play an important role in analyzing and responding to disease outbreaks, and COVID-19 has accelerated such uses. Livnat, Rhyne and Samore [5] describe Epinome, a visual tool that allows users to create and replay simulation scenarios, and investigate an ongoing outbreak with several different visualization tools. Lopez et al [6] describe work on the use of data analytics in predicting an influenza outbreak in India. Parwez, Abulaish, and Jahiruddin [7] performed a disease surveillance analysis by correlating reported disease cases with social network accounts of outbreaks.

Significant work has been done on analyzing COVID-19 data. Lueng et al [8] provide a description of a novel scheme for data visualization and general visualization tools used to visualize confirmed cases of COVID-19. Shang et al [9] describe a system used to analyze spatial COVID-19 data and Chen et al [10] describe a system to analyze temporal data regarding the pandemic. Podder and Podder [11] developed a mobile app that they claim can bring data analytics capabilities to end users. Nimpattanavong et al [12] analyzed data regarding impacts of COVID-19 on air traffic. The above is a sampling of major work on COVID-19 and clearly not exhaustive.

3. THE CURRENT STUDY

The current work comprises a case study in maintaining a dashboard that provides both summarizations of historical data on the COVID pandemic and on projections of measures including projected vaccinations. The following sections contain descriptions of the data sources used, the dashboard platform, and the means through which data was retrieved, processed, and presented.

[1] The authors wish to acknowledge and thank our peer editor, Dr. Amitabh Mishra.

Cases by Age Group:

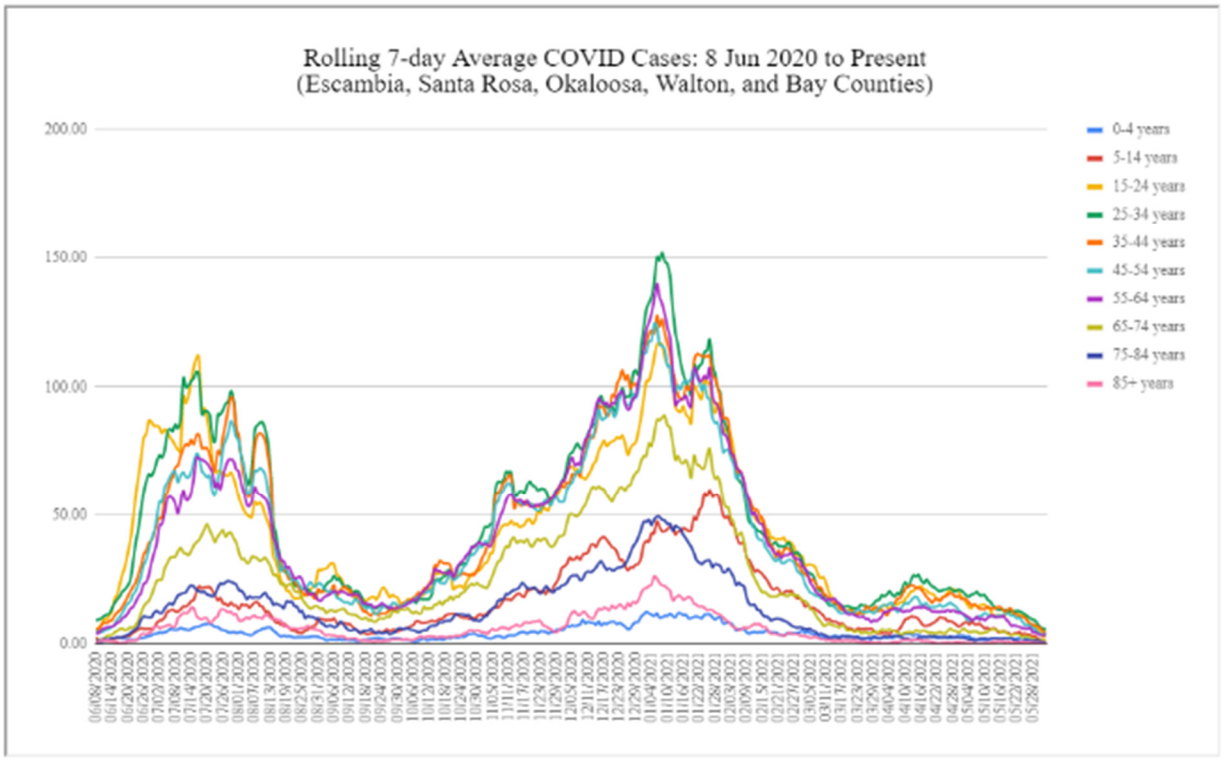


Figure 1. An example graphic presented in the dashboard.

3.1 Data Sources

Originally, data came from the Florida Department of Health (FL DOH) COVID19 Case Line database web site. Data in that repository was filtered by date and county, and by selecting desired output columns. Output of this process was a `.csv` file that was imported into an Excel workbook. The pivot tables in the workbook updated from the original source data. Seven-day average calculations were based on those pivot tables, and the charts based on those data.

3.2 Platform

This study used Google Sheets and Google Apps Scripts for the back-end data retrieval, data processing, and chart creation, and Github Pages to display the front-end dashboard. All of these services are free but have some limitations. Google Sheets has a 5 million cell limit for a workbook, and Google Apps Script had a 300s time limit for a script to run. Neither of these was a problem given the 5-county scope of the study, but both limits were exceeded in an expanded, state-wide study described later.

3.3 Data Retrieval

Data was retrieved directly from the FL DOH COVID19 Case Line database. A Google Apps Script used the current date to craft a query to the database that would request the desired columns, date ranges, and counties. The desired columns were: County, Age, Age Group, Number Hospitalized, Died, and EventDate. The FL DOH site responded with 2,000 results at a time in JSON format, which included a flag that indicated there were more results available. Checking the flag state allowed multiple loads to automate retrieval of the entire dataset in a single operation. A Google Apps Script paged through the responses and assembled them into a JSON object which was passed to an `ImportJSON` function [16]. This function updated a "Download" sheet in the Google Sheets workbook.

3.4 Data Processing

The main "Data" sheet in the workbook was a mirror of the six columns in the "Download" sheet, with two additional columns. From the "Data" sheet, additional sheets were used to host pivot tables created from the "Data" sheet for each desired dimension and measure combination (e.g. "Cases by County"), as well as to perform additional processing of the pivot table data. The created pivot tables provided counts of the desired measure (cases, hospitalization, or deaths) by date on each row, and either County or Age Group dimensions along each column.

Additional processing was performed to ensure there were no missing dates that could occur if there were zero occurrences of a chosen measure (such as Hospitalizations or Deaths) on a date in the date range of the pivot table. Missing dates were inserted into the date range with zero counts. At the time of the study, Google Sheets pivot tables did not natively support a "running average" calculation, so additional columns were used to calculate this measure for each pivot table date and dimension.

3.5 Presentation

The dashboard itself was hosted using Github Pages. Because the focus of the study was on the automated acquisition and processing of the data, a simple front-end was desirable. Thus the Github Page was made to link to the shared iframes of the charts that were created and updated in Google Sheets.

As the charts in the Google Sheet changed, the dashboard reflected the change the next time the page was loaded. Because of the amount of data the dashboard presented, loading times could approach 30 seconds, so a simple "Page is loading" dialogue was added. Figure 1 contains an example view of a typical dashboard graphic.

4. RESULTS

During this study, the amount of raw data for the 5-county local area increased from 1,563 rows in June 2020 to 102,718 rows in June, 2021. Data were checked hourly using a Google Apps Script, and when changes occurred, the Google Sheet was updated. This update typically occurred between 1-3pm US Central Time. After solving a "cascading dependency" issue where the pivot tables and charts would attempt to update during the data refresh, the update typically completed in 60 to 80 seconds.

The raw data were then summarized using pivot tables with dimensions of date, age group, school age group, and county, and with measures of cases, hospitalizations, and deaths. The pivot tables provided a sum of each measure for each day.

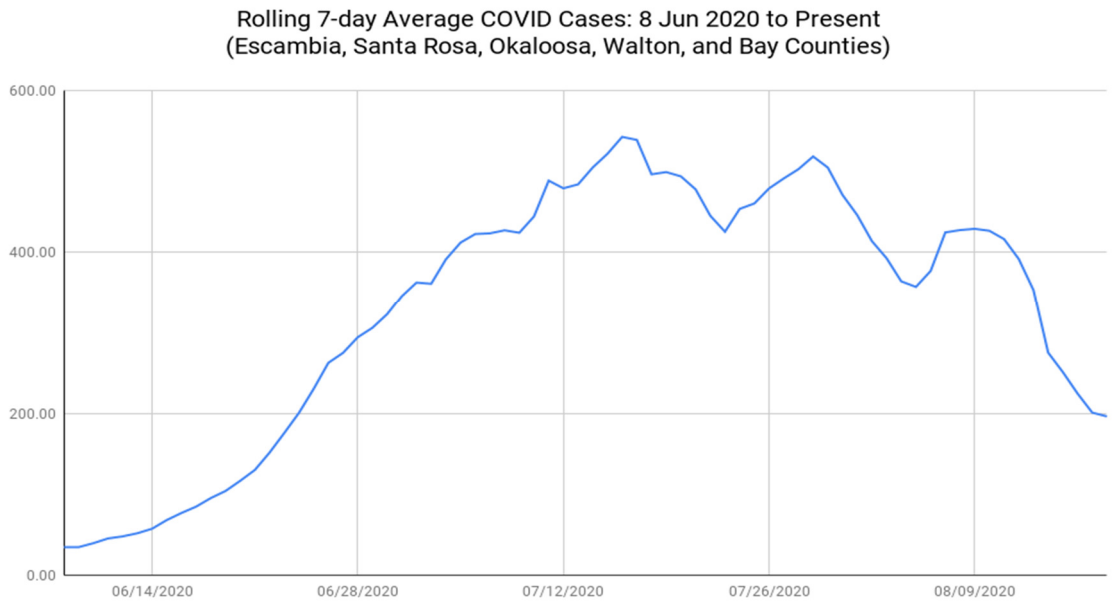
5. DISCUSSION

Two significant problems gave impetus to this work. One problem was that official reporting channels often presented data in a manner that lacked context, and that seemed to focus on alarming interpretations of the data. For example, on August 18, 2020 a local news site [18] published an article with the headline "Panhandle listed as hotspot by CDC, record deaths reported in Okaloosa County".

Both of those statements were factually correct, but ignored the fact that the referenced CDC report was for 1 June-15 July (a month prior), and that "both case counts and positivity rates for COVID-19 have shown a downward trend during most of this past two-week period" [18].

As seen in Figure 2, the dashboard created in the current study presented charts from that same date that conveyed the current state of the pandemic in an easier-to-understand context, and showed that local area cases had already decreased substantially from their July peaks.

A second problem was the need to report at the county level rather than at the state or national level. That was especially true in a state like Florida, with a variety of population densities and demographics. The charts were selected to give a snapshot view of how the pandemic was progressing in the local area. The current work has been ongoing since August 2020 with the publication of weekly snapshots of COVID data for a 5-county local area on a Facebook page.



:

Figure 2. A more nuanced depiction of case trends in the dashboard.

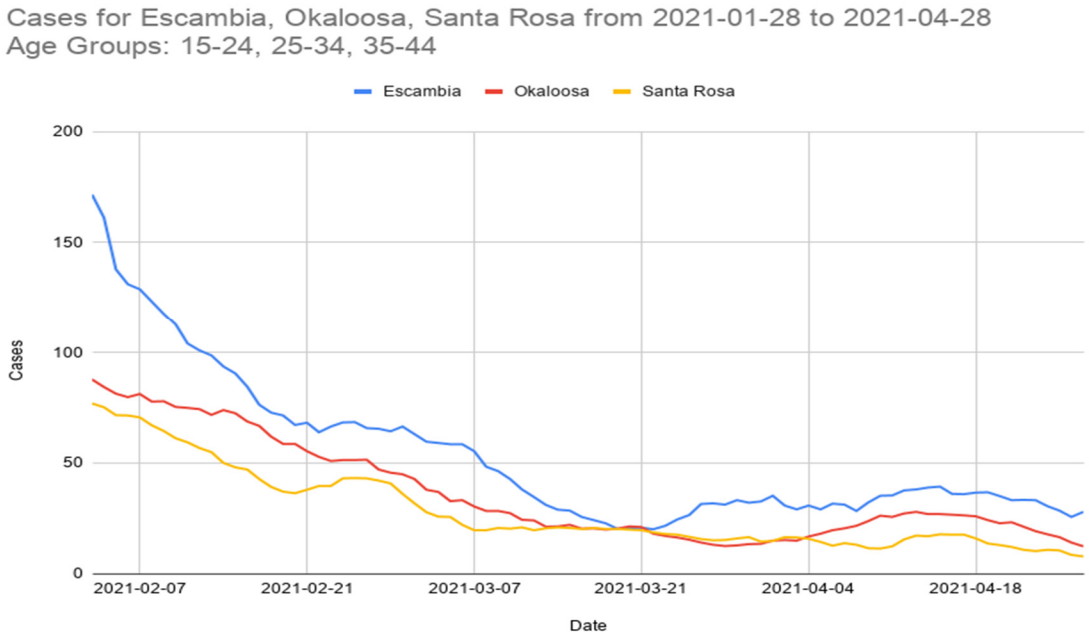


Figure 3. A Custom Chart created in a later phase of the study.

With the success of the initial 5-county dashboard, the decision was made to extend the dashboard to include the entire state of Florida. A custom chart Google Form was implemented, which drove a separate Google Sheet/Google Apps Script instance to get the data, create the chart, and email it to the requester. A sample result from that effort is illustrated in Figure 3.

Summarizing state-wide data presented a very different challenge. Given the 5 million cell limit for a Google Sheets workbook, the limit for the current dashboard was around 620,000 cases. That limit is based on 8 columns of data in the main "Data" sheet, with some cells reserved for the pivot tables and 7-day average formulas in the rest of the workbook.

As of 9 Jul 2021, the state of FL had reported 2,361,360 cases [19] so to expand the dashboard statewide another solution was needed. Microsoft Excel, OSX Numbers, and OpenOffice Calc all have 1 million row limits. Gnumeric has a 16 million row limit, but lacks support for pivot tables or SQL queries against the dataset, both of which Google Sheets supports. The statewide dataset was imported into Gnumeric, but that solution was unusable (very slow to respond) on two commodity computers with 16 and 48 GB of RAM.

Tableau Public was tested as a means to implement both custom charts and to process the statewide dataset, and afforded immediate success. The Tableau Public project is easily shareable, so users could filter by date range, county, age group and measure in any combination. Tableau Public offered both the scalability to a statewide (or national) data set as well as the ability for users to customize charts.

A final lesson learned was from something the project did not implement but should have: a daily backup schedule for both the raw data and the charts as presented on the dashboard. Backups would have allowed retrospective looks at when fields in the data had changed. One use of this information would have been to develop a rule stating when it would be reasonable to no longer refresh records. It would also have allowed the inherent delays present when working with large datasets to be noted. This addition would have enabled the dashboard to do a better job of presenting the time elapsed between cases rising, hospitalizations and deaths rising.

6. CONCLUSIONS

This work supports the claim that creating and sustaining a dashboard summarization of large datasets can be achieved with low cost and open source tools. The current work illustrates the fact that issues of scope and scale loom large in such work. Properly focused, relatively small scale data analytics projects might quickly find great utility as smaller organizations try to make sense of relevant internal or external data. Such systems readily become susceptible to "mission creep" and the need to handle larger, changing datasets. However, tools that work well for smaller datasets might not scale, necessitating hard decisions regarding migrations and potential changes to or loss of existing features. Separately, if the data being summarized is external to the organization as is most likely to be the case in a work like the one described in this article, data analysts are at the mercy of the data publishing organization. Despite these potential difficulties, right-sizing smaller-scale data analytics initiatives holds great potential to lend great value to small organizations.

7. ACKNOWLEDGEMENTS

The authors wish to acknowledge and thank our non-anonymous, peer reviewer, Dr. Steve Bitner.

8. REFERENCES

[1] A. Yamada and M. Peran, **Governance framework for enterprise analytics and data.** 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3623-3631, doi: 10.1109/BigData.2017.8258356.

[2] N.W. Grady, J.A. Payne, H. Parker. **Agile big data analytics: AnalyticsOps for data science.** 2017 IEEE International Conference on Big Data (Big Data). DOI: 10.1109/BigData.2017.8258187.

[3] A. Londhe and P. P. Rao, **Platforms for big data analytics: Trend towards hybrid era.** 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 3235-3238, doi: 10.1109/ICECDS.2017.8390056.

[4] L. Ehrlinger, T. Grubinger, B. Varga, M. Pichler, T. Natschläger, J. Zeindl, **Treating Missing Data in Industrial Data Analytics.** 2018 Thirteenth International Conference on Digital Information Management (ICDIM), 2018, pp. 148-155, doi: 10.1109/ICDIM.2018.8846984.

[5] Y. Livnat, T. Rhyne, M. Samore, **Epinome: A Visual-Analytics Workbench for Epidemiology Data.** in IEEE Computer Graphics and Applications, vol. 32, no. 2, pp. 89-95, March-April 2012, doi: 10.1109/MCG.2012.31.

[6] D. Lopez, M. Gunasekaran, B. S. Murugan, H. Kaur, K. M. Abbas, **Spatial big data analytics of influenza epidemic in Vellore, India.** 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 19-24, doi: 10.1109/BigData.2014.7004422.

[7] M. A. Parwez, M. Abulaish and J. Jahiruddin, **A Social Media Time-Series Data Analytics Approach for Digital Epidemiology.** 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2020, pp. 852-859, doi: 10.1109/WIIAT50758.2020.00131.

[8] C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, Y. Wen, A. Cuzzocrea, **Big Data Visualization and Visual Analytics of COVID-19 Data.** 2020 24th International Conference Information Visualisation (IV), 2020, pp. 415-420, doi: 10.1109/IV51561.2020.00073.

[9] S. Shang, C. K. Leung, Y. Chen, A. G. M. Pazdor, **Spatial Data Science of COVID-19 Data.** 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2020, pp. 1370-1375, doi: 10.1109/HPCC-SmartCity-DSS50907.2020.00177.

[10] Y. Chen, C. K. Leung, S. Shang, Q. Wen, **Temporal Data Analytics on COVID-19 Data with Ubiquitous Computing.** 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 2020, pp. 958-965, doi: 10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00146.

[11] A. Poddar, M. Poddar, Covid-19 Data **Visualization and Data Analytics with a Smart Standalone Mobile Application.** 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342143.

[12] C. Nimpattanavong, P. Khamlae, W. Choensawat, K. Sookhanaphibarn, **Flight Traffic Visual Analytics during COVID-19.** 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), 2020, pp. 215-217, doi: 10.1109/GCCE50665.2020.9291896.

[13] G.P. Duggan, D. Zimmerle, S. Upadhyay. **Big Data Analytics for Power Distribution Systems using AMI and Open Source Tools.** In 2020 IEEE/PES Transmission and Distribution Conference and Exposition (T&D) (pp. 1-5). IEEE, doi: 10.1109/TD39804.2020.9299952.

[14] **Florida COVID-19 Data**. Accessed on: Jul 26, 2021. [Online] Available: https://experience.arcgis.com/experience/d2726d6c01c4486181fec2d4373b01fa/page/page_0

[15] Florida Department of Health Open Data. **Florida COVID-19 Case Line Data,** Accessed on: Jul 6, 2020. [Online] Available: https://open-fdoh.hub.arcgis.com/datasets/florida-covid19-case-line-data/about

[16] GitHub. **bradjasper/ImportJSON: Import JSON into Google Sheets** Accessed on: Jul 26, 2021. [Online] Available: https://github.com/bradjasper/ImportJSON

[17] **NW FL COVID-19 Tracking Dashboard,** Accessed on: Jul 26, 2021. [Online] Available: https://ssmessia.github.io/

[18] W. Victora, **Panhandle listed as hotspot by CDC, record deaths reported in Okaloosa County.** Northwest Florida Daily News Aug 18, 2020, Accessed on: Jul 12, 2021. [Online] Available: https://www.nwfdailynews.com/story/news/2020/08/18/florida-panhandle-listed-hotspot-cdc-okaloosa-record-deaths/3390270001/

[19] **COVID-19 Weekly Situation Report: State Overview.** Accessed on: Jul 14, 2021. [Online] Available: http://ww11.doh.state.fl.us/comm/_partners/covid19_report_archive/covid19-data/covid19_data_20210709.pdf