Improving Efficiency in Large-Scale Data Processing and Analysis Using Public Datasets and Cloud Computing

A Professional Readiness Experiential Program (PREP) Project Effort

----- ***Authors / Student Project Team Members*** -----

Sri Avula is a student at George Mason University graduating with a bachelor's degree in Management Information Systems. Sri is looking to get a job related to Data Analytics post-graduation in December 2025. Sri enjoys nature, and motorcycles as well as cooking in his spare time.

Thanh Quach is a student at George Mason University graduating with a bachelor's degree in Business Analytics. He has experience in data research, data cleaning, data visualization, and building data pipelines using RStudio and Python. He is currently seeking job opportunities after his graduation in May 2026. He hopes to travel around the world after retirement.

Amena Zaini is a student at George Mason University graduating with a bachelor's degree in Business Analytics and a minor in Management Information Systems. She is interested in careers within data, financial, and operations analytics, and enjoys applying analytical tools to solve real-world problems. Through her internship experience, she gained hands-on experience working with sensitive data and secure systems that helped grow her analytical and technical skills.

Denis Bykov is a student at George Mason University graduating with a bachelor's degree in Management Information Systems, and he's really focused on technology, cybersecurity, and improving his skills through hands-on projects and internships. He is hardworking, detail-oriented, and always looking for ways to grow personally and professionally.

Aaliyah Alva is a student at George Mason University graduating with a bachelor's degree in Management Information Systems. She brings practical experience from her AWS Cloud Support Associate internship, where she worked in the cloud environment and earned two certifications. After graduation, Aaliyah hopes to continue her career in AWS as a Cloud Support Engineer or Solutions Architect.

----- ***Industry Participant / Mentor*** -----	----- ***Faculty Member*** -----
John Blair - *Director, National Business Investment* ***Stephen Tarditi*** - *Director, Market Intelligence* ***Richard Smith*** - *AWS Technical Consultant* ***John Hoeveler*** - *National Business Investment*	**Brian K. Ngac, PhD** FWI Corporate Partner Faculty Fellow Instructional Faculty & Dean's Teaching Fellow
FairFax County Economic Development Authority (FCEDA)	George Mason University's Costello College of Business

Interested in being an Industry Participant and or PREP Sponsor? Please reach out to bngac@gmu.edu, Thanks!

Introduction

Fairfax County Economic Development Authority (FCEDA) helps businesses start, expand or relocate in Fairfax NOVA. Chartered by the Commonwealth of Virginia, FCEDA is led by a commission of county business leaders and funded by the Fairfax County government to promote Fairfax NOVA as one of the world's best business locations.

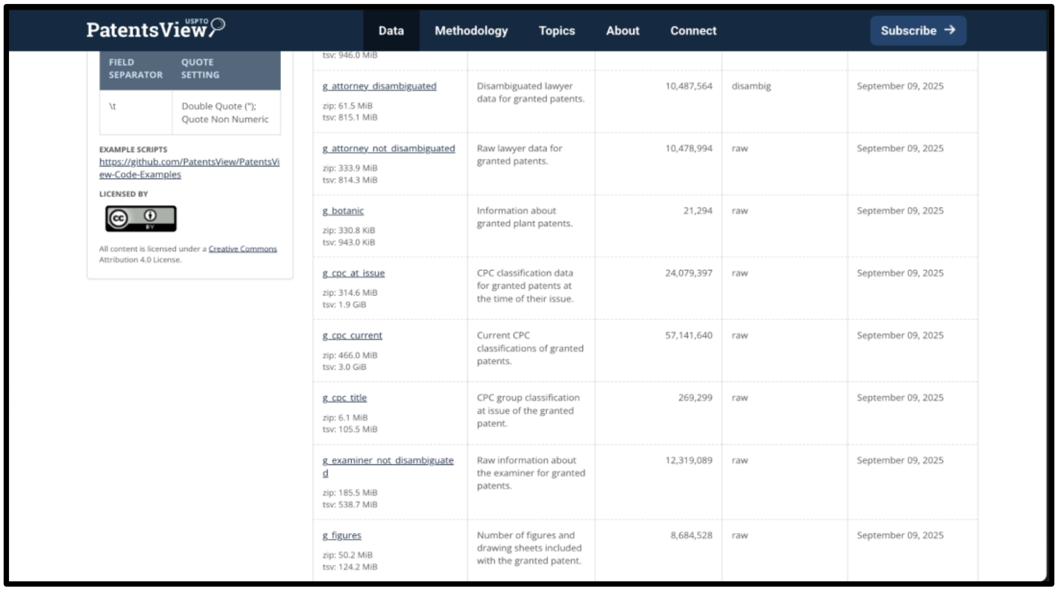
FCEDA has three business investment divisions – National, International and Business Diversity and Entrepreneurship – that work with companies interested in starting, expanding and relocating businesses to Fairfax NOVA. FCEDA also has market intelligence, real estate services and communications divisions.

Business Challenge

FCEDA needs assistance in developing a solution that will quickly clean, enhance and visualize large datasets for analytical and outreach purposes. FCEDA will provide a data set from the United States Patent Trademark Office as the test case. Identifying new business at an early stage is an important goal for FCEDA, and they would like to develop a solution that allows them to analyze these large data files from Public Datasets. Cleaning these data files is a time consuming process that isn't feasible for FCEDA to complete manually.

The Business Investment and Market Intelligence teams are looking for a quicker solution that would allow them to clean, filter, and merge together individual datasets into a clean file that would allow them create visualizations from. These visualizations can be implemented on a dashboard or other tools to allow them to present in a neat and easy to understand format. Currently, staff members depend on lengthy manual processes for data collection, cleaning and reporting.

The image below shows where the raw patent data is available online. There's millions of rows in a TSV File format, which is a harder format to work with. The size of the files also pose issues as it will take more time to process, especially manually.



FIELD SEPARATOR	QUOTE SETTING
\t	Double Quote ("); Quote Non Numeric

EXAMPLE SCRIPTS
https://github.com/PatentsView/PatentsView-Code-Examples

LICENSED BY

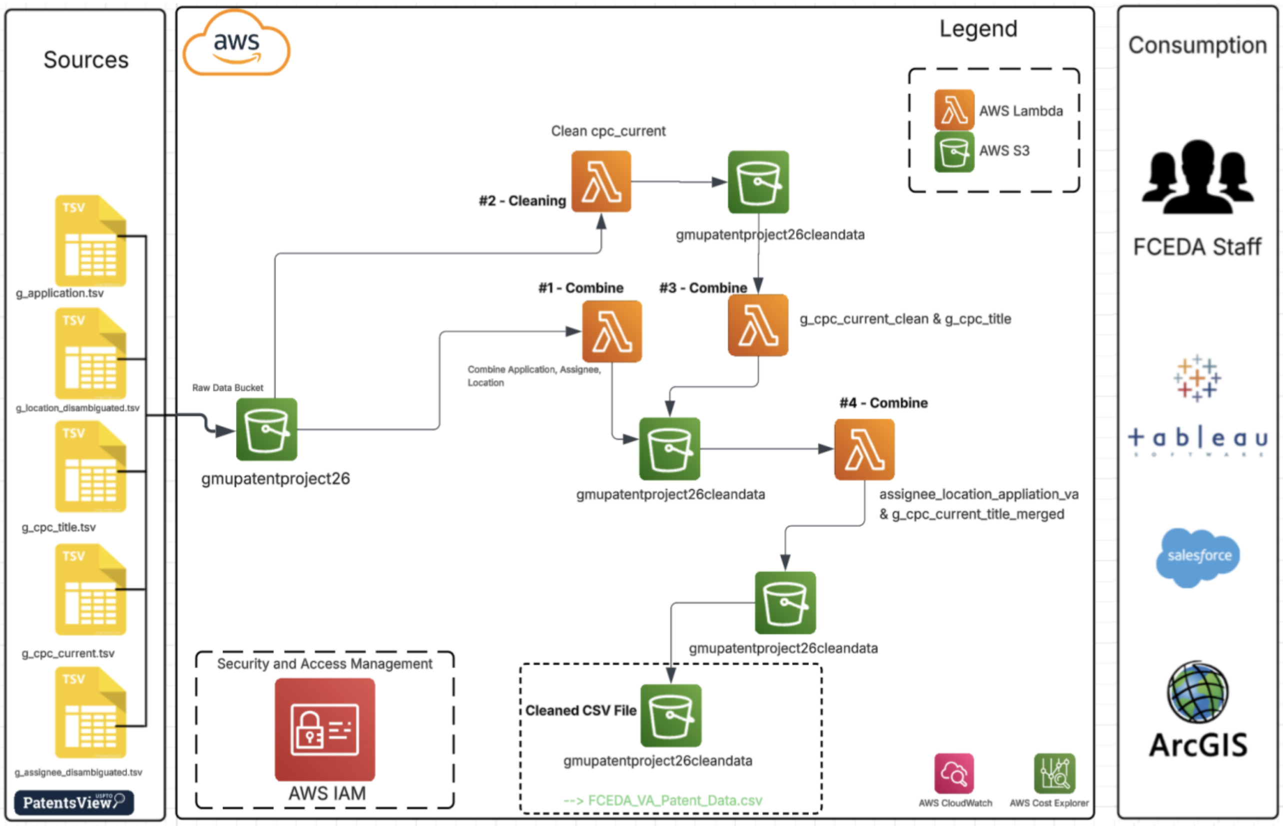
All content is licensed under a Creative Commons Attribution 4.0 License.

tsv: 946.0 MiB				
g_attorney_disambiguated zip: 61.5 MiB tsv: 815.1 MiB	Disambiguated lawyer data for granted patents.	10,487,564	disambig	September 09, 2025
g_attorney_not_disambiguated zip: 333.9 MiB tsv: 814.3 MiB	Raw lawyer data for granted patents.	10,478,994	raw	September 09, 2025
g_botanic zip: 330.8 KiB tsv: 943.0 KiB	Information about granted plant patents.	21,294	raw	September 09, 2025
g_cpc_at_issue zip: 314.6 MiB tsv: 1.9 GiB	CPC classification data for granted patents at the time of their issue.	24,079,397	raw	September 09, 2025
g_cpc_current zip: 466.0 MiB tsv: 3.0 GiB	Current CPC classifications of granted patents.	57,141,640	raw	September 09, 2025
g_cpc_title zip: 6.1 MiB tsv: 105.5 MiB	CPC group classification at issue of the granted patent.	269,299	raw	September 09, 2025
g_examiner_not_disambiguated zip: 185.5 MiB tsv: 538.7 MiB	Raw information about the examiner for granted patents.	12,319,089	raw	September 09, 2025
g_figures zip: 50.2 MiB tsv: 124.2 MiB	Number of figures and drawing sheets included with the granted patent.	8,684,528	raw	September 09, 2025

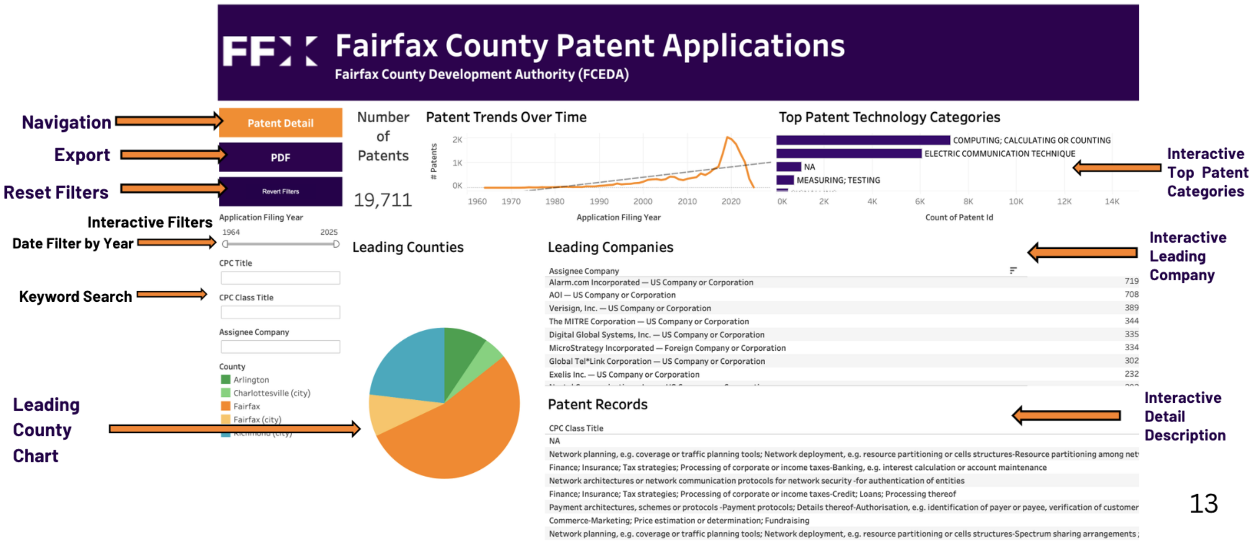
Activities Done to Address the Business Challenge

USPTO Patents Data was combined, cleaned, and filtered into a neat data set through an automated AWS ETL Pipeline Solution. This allows for easy data analysis by FCEDA Staff in the future as Patents Data updates. This data can be displayed on a Tableau Dashboard for quick and easy data visualization. A Tableau Dashboard was also created that was tailored to FCEDA's preferences, which allows them to analyze the Patent Data for factors they find important. Both of these were created to be as easy to use as possible by FCEDA staff, and even users with minimal technical experience will be able to use these. This also requires very little time to set everything up, even as patent data updates in the future. FCEDA would only spend about five minutes setting everything up for analysis. The computing of the data, and cleaning would take 20 - 40 minutes, but this can be done in the background, while staff complete other tasks.

The image below goes over our ETL Pipeline. It summarizes our automated solution and how it works on the cloud. This Solution allows FCEDA to easily access patent data, and perform analysis on a frequent basis.


Sources
g_application.tsv
g_location_disambiguated.tsv
g_cpc_title.tsv
g_cpc_current.tsv
g_assignee_disambiguated.tsv
PatentsView
aws
Legend
AWS Lambda
AWS S3
Clean cpc_current
#2 - Cleaning
gmupatentproject26cleandata
#1 - Combine
#3 - Combine
g_cpc_current_clean & g_cpc_title
Combine Application, Assignee, Location
Raw Data Bucket
gmupatentproject26
#4 - Combine
gmupatentproject26cleandata
assignee_location_appliation_va & g_cpc_current_title_merged
gmupatentproject26cleandata
Security and Access Management
AWS IAM
Cleaned CSV File
gmupatentproject26cleandata
--> FCEDA_VA_Patent_Data.csv
AWS CloudWatch
AWS Cost Explorer
Consumption
FCEDA Staff
tableau
salesforce
ArcGIS

The image below shows our final Tableau Dashboard solution that we constructed for FCEDA to use. They can use this dashboard solution to perform analysis on the patent data in an easy to understand format. They can also easily export this for presentation to higher executives.



Leading Companies

Assignee Company	
Alarm.com Incorporated — US Company or Corporation	719
AOI — US Company or Corporation	708
Verisign, Inc. — US Company or Corporation	389
The MITRE Corporation — US Company or Corporation	344
Digital Global Systems, Inc. — US Company or Corporation	335
MicroStrategy Incorporated — Foreign Company or Corporation	334
Global Tel*Link Corporation — US Company or Corporation	302
Exelis Inc. — US Company or Corporation	232

Patent Records

CPC Class Title
NA
Network planning, e.g. coverage or traffic planning tools; Network deployment, e.g. resource partitioning or cells structures-Resource partitioning among net
Finance; Insurance; Tax strategies; Processing of corporate or income taxes-Banking, e.g. interest calculation or account maintenance
Network architectures or network communication protocols for network security -for authentication of entities
Finance; Insurance; Tax strategies; Processing of corporate or income taxes-Credit; Loans; Processing thereof
Payment architectures, schemes or protocols -Payment protocols; Details thereof-Authorisation, e.g. identification of payer or payee, verification of customer
Commerce-Marketing; Price estimation or determination; Fundraising
Network planning, e.g. coverage or traffic planning tools; Network deployment, e.g. resource partitioning or cells structures-Spectrum sharing arrangements ;

Results & The Positive Impact

This neat solution allows for FCEDA Staff to analyze the patent data on a much more frequent basis, and with minimal manual labor. This can be used to analyze and present the patent data as it updates on a frequent basis. They can use this information to their specific needs as necessary based on their goals mentioned regarding identifying new businesses at an early stage. This is a very important tool for FCEDA that will allow them to conduct key research on a new area that previously they were unable to conduct on a frequent basis. This key information can assist them in the future of checking in on businesses as needed, to see if they can provide support.

Conclusion

This solution constructed by our PREP Team allows FCEDA Staff to easily upload new patent data for cleaning into the AWS ETL Pipeline solution. This then provides them a clean CSV File output that is cleaned, combined, and filtered based on FCEDA's preferences. This file can then be used to perform analysis on Tableau, ArcGIS and Salesforce. For Tableau we constructed a dashboard tailored to FCEDA's preferences, that can be easily used by FCEDA Staff.

PREP Student Reflection

It was a wonderful opportunity to work with FCEDA throughout the semester; the team was very kind and helpful throughout. We want to thank Stephen, John and Richard for their continuous support throughout the semester. We were able to learn in an agile environment and gained important experience with AWS services like S3, Lambda, and Cloudwatch. We were also able to adapt to using various tools within Tableau, where we had to learn the skills required to create a clean dashboard that FCEDA could use in a professional environment. We also learned how to work with very large unfamiliar data files, and create a process to clean, combine and filter those down to what FCEDA wanted. This was an important project for our team, that we worked very hard on throughout the semester and were able to hand off a final deliverable to the FCEDA Team. In the end we analyzed millions of rows of raw data, which we then cleaned, combined, filtered down to FCEDA's specific preferences, we then automated this process so FCEDA can replicate with ease in the future, lastly we constructed a Tableau Dashboard based on FCEDA's preferences that they can use to conduct analysis in the future as necessary.