Measuring the Impact of the COVID-19 Pandemic Using Historical Companies Websites in European Union Countries

Dirga Imam Gozali Sumantri

A thesis present for
Master of Spatial Transport Regional Environmental Economics

University: Vrije Universiteit Amsterdam
Supervisor: Michael D. König
Student Number: 2731982
Date: 10-08-2022

Abstract

The spread of COVID-19 causes a disruption in the production, demand, and supply chain of the economy. These necessitates the companies to new working environment and business operation which often reflected by company policies. Company websites are introduced as the new data source to measure the impact of COVID-19 on companies. Processing a large amount of company website data requires a big data framework and cloud computing processing. We are able to crawl data from around four million website pages which represent approximately fifty-three thousand companies in European Union. Also, we are able to build panel data with quarterly temporal resolution from the beginning until the new normal condition of COVID-19. The data from this research can be used for further analysis related to the impact of COVID-19 on companies.

Acknowledgement

I would like to express my deepest gratitude to my thesis supervisor, associate professor Michael D. König for giving me the opportunity with this challenging process and topic.

Special thanks to my colleague M Yudha Pratama for always supporting me and giving positive peer effects during my study.

To my wife (Putri) and my son (Raik).

Table of Contents

1. Introduction

Since the first case of COVID-19, it has taken more than years to return on the no-COVID-19 restriction policy in most European countries. Back to normal condition and settling the COVID-19 as a common disease, we went through an economic shock period caused by the disease. The spread of COVID-19 is transmitted by direct contact between humans, which disrupts how people communicate and commute. During the pandemic period, COVID-19 took more than half a billion cases and caused more than six million deaths (*Worlodometer*, 2022).

COVID-19 causes a disruption in the production, demand, and supply chain of the economy (Maital & Barzani, 2020). Measured by the performance, firms are disrupted significantly by COVID 19 (Shen et al., 2020). Companies issue several policies to adapt COVID-19 as the new normal condition: working from home and employee layoffs. These companies' actions will lead to lower employment and consumer spending, lowering the demand side. In addition to reducing production, mobility restrictions also disturb the companies' production supply chain. More so for the multinational companies which require transactions across borders.

Currently, the rise of digital creates the information transaction rapidly. As the spread of information can be captured within seconds or minutes, firms can respond to any relevant topic with the market condition. A company website is one tool to communicate information representing a resource for organizations, such as corporate information or policies. The information of companies related to specific topics is valuable to observe the company's behavioural change and might change the movement of the economy and vice versa.

By observing the response of firms to COVID-19 in the digital era, several contributions are made to this study. First, we want to unlock the new potential of revealed preference data sources. The internet produces petabytes of data every day. We crawl the data provided by Common Crawl, a non-profit organization that stores the entire website archives worldwide. We create an end-to-end scheme for the big data process using cloud computing. Cloud computing is utilized for almost the whole crawling process. By targeting the company's website, we want to extract text information related to COVID-19.

This entire crawling process has never been well documented before. We want a straightforward process where the procedure is replicable without having high programming skills. Several technology stacks use no-code platforms, easily implemented by other researchers. Also, the implementation of cloud computing can be operated on any computer without worrying about the computer capability, whereas a single desktop or computer has computing limitations.

Second, we want to measure how companies respond to COVID-19 from the beginning until the end of the pandemic. We build panel data from the revealed preference data and see to what extent this data is valuable for many studies. The WARC files are identified to find the keywords related to the COVID-19 pandemic. Then, we implement sentiment analysis on the COVID-19 information to measure the sentiment analysis of companies during the COVID-19 pandemic.

However, previous studies examining the effects of COVID-19 often rely on survey data by which the companies covered are very limited. For example, the Ifo Business Survey employed by (Buchheim et al., 2022) consists of around six thousand German companies. Even though it can be argued that the data is representative enough for the economy, we cannot conduct research on a larger scope, such as the effects of COVID-19 in Europe. It is important for a study with a larger scope because the pandemic affects across nations (Maital & Barzani, 2020)

We managed to run the data crawling, where this can be a new potential of revealed preference data for many topics. A technical scheme is produced to obtain the web archive data with a specific target. We implement the cloud computing service provided by Amazon Web Service (AWS) for many tasks such as data queries, big data preparation and visualization, running the crawl script, and storing the table result. The whole process requires a basic understanding of SQL and Python programming languages.

We find around 172,000 has a digital footprint on Common Crawl, where 53,698 companies put the COVID-19 keywords as the URL target of the website. Panel data has been created within around two years with nine different dates. A corpus dataset has been established and can be used for further analysis, such as sentiment and polarity analysis of companies during the COVID-19 pandemic.

The research is organized on the following structure:

- The introduction part provides the motivation and research gap of the study
- The literature studies part provides the relevant concept and theory of the study
- The methodology part provides the initial data and strategy used in the study
- The result part summarizes the study result and analysis.
- The conclusion and limitation part describes the learning points, limitations and improvements for future study

2. Literature Study

Firms are the part of crucial entities in economics which produce millions of goods (Nicholson & Snyder, 2008). Individuals are involved in consumption as consumers and production as labourers (Nicholson & Snyder, 2008). The number of goods produced by firms will behave the market condition (Mason, 1939). In the digital era, firms build a digital twin between the real-world and digital world as the technology factor, which neither increases production nor efficiency (Trauer et al., 2021). As the spread of information can be captured within seconds or minutes, the information spillover of the market condition can be responded to favourably by firms (Benveniste et al., 2003). The ability to catch the issue which disrupts the company's activities and be responsive will mitigate any unpredictable cost spikes (Deloitte, 2020). In this thesis, we are able to obtain the company's information as they respond to specific topics.

The rise of the World Wide Web (www) or website technology and framework has become a game-changer in how people make decisions, study, and communicate (Constantinides & Fountain, 2008). The number of registered domains has reached more than 1.9 billion worldwide (*Internetlivestat*, 2022). The website substitute the way people and organization exchange information over the internet, including the presence of company or corporate websites (Heinze & Hu, 2006). A company website, as one of the website categories based on its function, is defined as all web pages owned and operated by a company using the representative domain name (*Lawinsider*, 2022). The website enables companies to publicly communicate information that represents a resource for organizations, such as corporate information, social issues, and employment opportunities (Robbins & Stylianou, 2003).

Some empirical works of literature have been performed by looking at various company websites that can achieve new findings from specific topics. (Daas & van der Doef, 2020) identifies the innovation by analyzing the related keyword to innovation from thousands of websites in the Netherlands. The results revealed that it is possible to determine if companies are technological innovative based on the text analysis on its website (Daas & van der Doef, 2020). Gamerschlag et al., 2011 measured the Corporate Social Responsibility (CSR) disclosure of German companies by analyzing text related to CSR from its company website. The website is considered the main channel of communication, which integrate CSR-related aspect to company publication which might be attached to financial or human capital reports and other media such as press release (Gamerschlag et al., 2011).

Contradictory to the real world, where all past events only become remembrance, the digital world produces historical information which is accessible and remains unchanged (El-Showk, 2018). All texts, images, videos, and other formats on the internet can be stored as archives (Niu, 2012). These archives become a new advantage but also a problem, where people will produce 1.7 Megabytes per second per capita in 2020 (Bulao, 2022). The Common Crawl, a non-profit organization which captures the web archives in the world periodically, stores petabytes of billions of websites for a single capture (CommonCrawl.org, 2022). This thesis crawls the information that is stored by Common Crawl via Amazon Public Dataset.

The term big data and cloud computing are often used concurrently. Cloud computing is related to the process of computing a massive amount of data in IT services providers (Goos et al., 2009). Big data often cannot be processed by a personal computer because its computing requires a large amount of computing power and large storage (Kaisler et al., 2013). Therefore, cloud computing main features are its large-scale computing capabilities and its scalability (Dillon et al., 2010). The elastic scalability feature is important because the computing power can be utilized on demand; that is, depending on the requirement, we can scale up the capacity of the data processing calls for it (Kaisler et al., 2013). Because of its convenience, cloud computing services have been adopted by both companies and private users (Dillon et al., 2010). In this thesis, we also utilize cloud computing by AWS to process the big data of the company websites. We will explain this more thoroughly in the Methodology section.

The processing of company websites requires text mining, in which the text is retrieved from information applications to be analyzed further (Aggarwal & Zhai, 2012). This text mining will generate a large amount of text data, which must be cleaned and structured for further analysis (Tang et al., 2016). One of the text analysis procedures to detect the views contained in a corpus (i.e. the text on a web page) is called sentiment analysis (Tang et al., 2016). In the case of a company website, the sentiments of that website could be analyzed through sentiment analysis by classifying the terms used in the website: usually can be classified as positive, negative, and neutral sentiment (Heerschop et al., 2011). This sentiment analysis could then be connected to the economic analysis based on our research questions.

COVID-19 has a significant impact to the entire individuals, firms, and society. Particularly, the pandemic causes a disruption in the production, demand, and supply chain of the economy (Maital & Barzani, 2020). Initially, the impact starts with the shock in production, in which workers must be limited to prevent the spread of the virus (Loayza & Pennings, 2020). The limitation to moving across the border also disrupts the supply chain. (Baldwin et al., 2020). Consumers will also reduce their spending due to lower employment (Chetty et al., 2020). Overall, the impact of COVID-19 is the contraction of both supply and demand in the global economy (Loayza & Pennings, 2020).

The uncertainty of market conditions due to external shocks such as COVID-19 will impact the disruption among firms (Shen et al., 2020). Firms' response to COVID-19 may be reflected by their policy or statement on their website. Sentiment analysis may be a proper approach to measure the sentiment of firms in response to COVID-19.

3. Methodology

In this chapter, we want elaborate on the data, methodology, and technology during this research. The research design is depicted in Figure 1. Diagram A depicts the process depending on the task for each step. Diagram B depicts the process depending on the cloud computing service which is used during the process. The diagram workflow is arranged to describe the whole picture of the study and will be elaborated on chapter 3 and chapter 4.

- Section 3.1.1 explains the process on A-1.
- Section 3.1.2 explains the process on A-2.
- Section 3.2.1 explains the process on A-3 and A-4.
- Section 3.2.2 explains the process on A-5, A-6, and A-7.
- Section 3.3.1 explains the technology on B1 and B2
- Section 3.3.2 explains the technology on B3 and B5
- Section 3.3.3 explains the technology on B8
- Section 3.3.4 explains the technology on B10

3.1 Data

3.1.1 Companies in EU Data

Companies' information is the main subject of this study. First, we need to find the companies' information in EU countries consisting of 27 countries. The data is obtained from Bureau van Dijk or ORBIS, which provides different types of company information. ORBIS claim they record nearly 400,000,000 companies across the globe. From the ORBIS database, we gather 594,213 companies, including important information such as company name, location, sector, financial information, and website address. The distribution of companies is depicted in Figure 2, and Figure 3 shows the share of companies across countries and sectors.

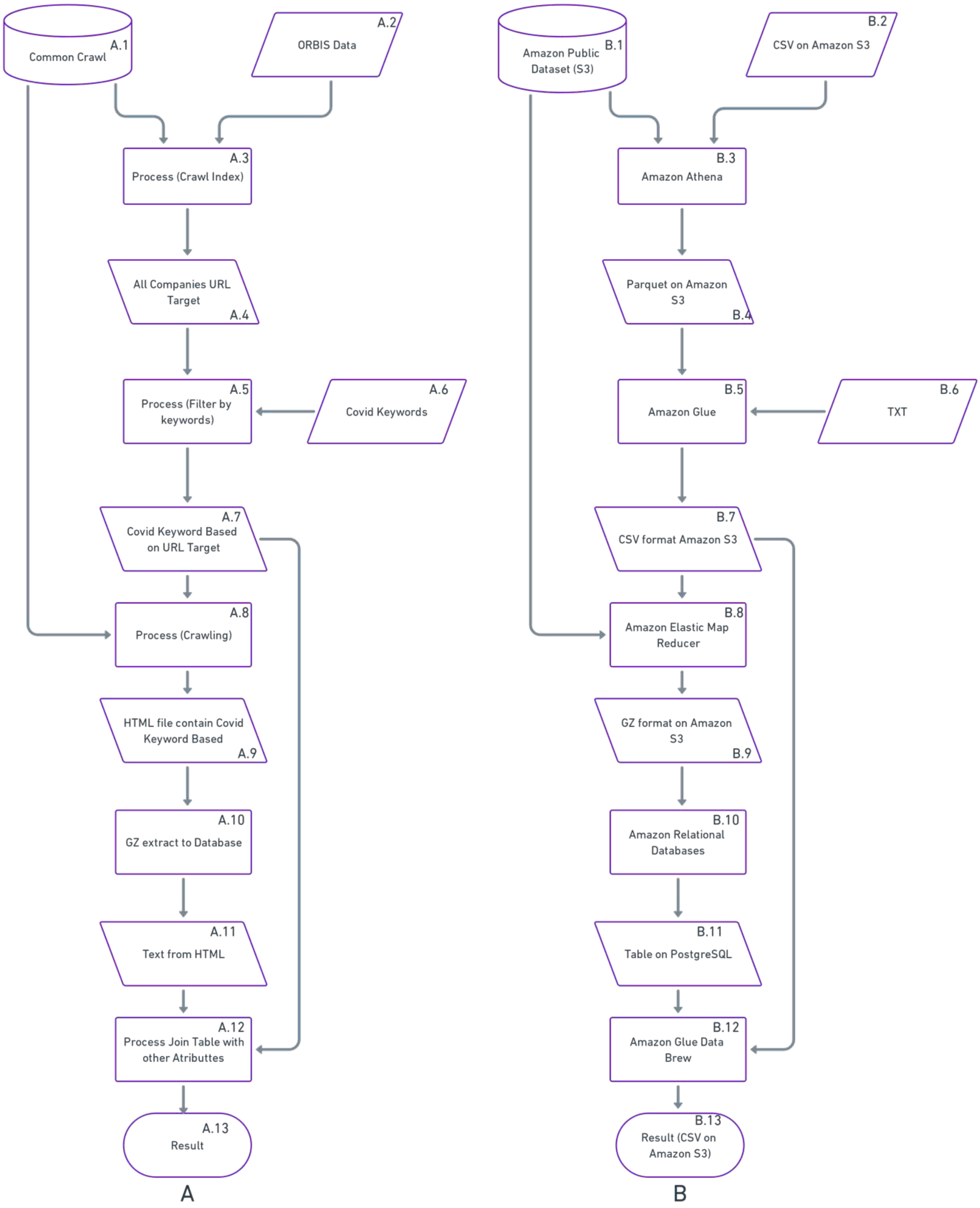


Figure 1. Research Design

Data cleaning must be performed to prevent any errors when processing the data. Companies' information which does not contain website addresses is removed. Without a website address, we are unable to locate the web archive. Even though this could introduce a selection bias for further analysis, such as causal inference, that analysis is beyond the scope of this thesis. A web archive is used as the data target where the text can be retrieved for analysis. The representative location is chosen randomly, so this does not highly affect our analysis. Several companies are multi-region companies which have representatives in several countries. To prevent duplication, we only pick one representative location. The company is a unique value and only represents one country. Furthermore, we adjust the website address format according to Common Crawl criteria, such as removing WWW or HTTP from the address. This process reduces the number of companies and leaves 298,863 companies.

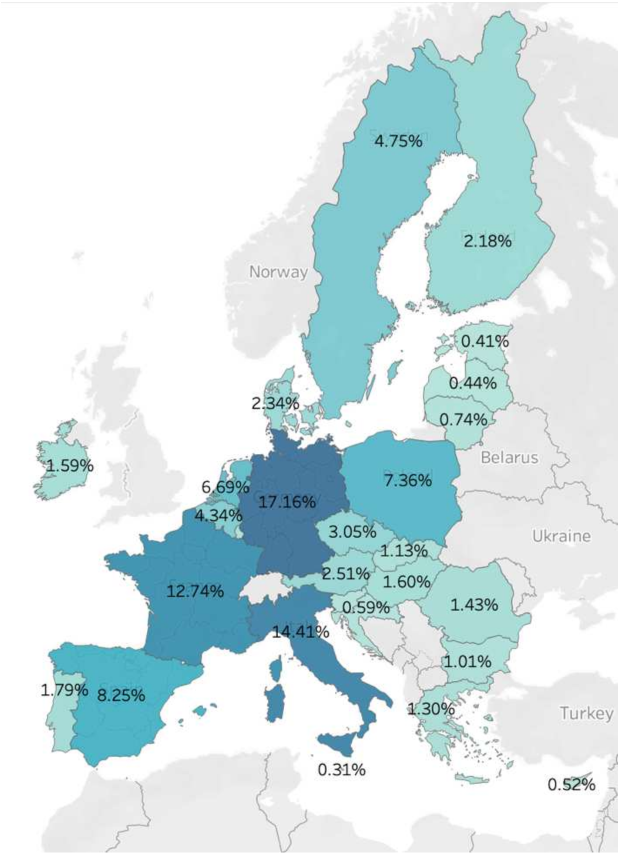


Figure 2. Share of Companies Across Countries

During this pre-processing step, we perform the data cleaning of the website address using Microsoft Excel Desktop. The amount of data on this step is properly processed on the desktop application. When the data processing can no longer be accommodated by desktop applications, we utilize cloud computing to speed up the process and avoid failure or crash. Cloud computing will be performed in several next steps.

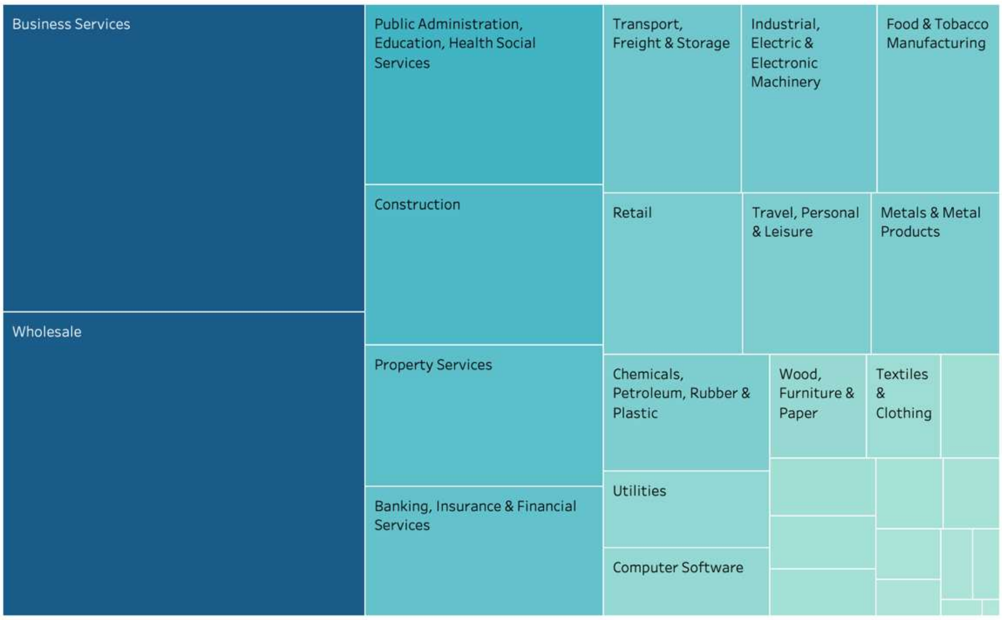


Figure 3. Share of Companies Across Sectors

3.1.2 Web Archives Data

We want to get the historical information of companies from web archives provided by Common Crawl. Common Crawl is a non-profit organization which builds and maintains an open repository of website crawl data. Common Crawl web archives are stored on Amazon Web Service (AWS) Open Data and can be accessed freely. For each crawling, the uncompressed data can reach more than 300 TiB. The data is crawled periodically where it contains billions of pages and petabytes of memory usage. Even if the data is free of charge, processing the huge amount of data requires high-performance computing which is costly. Therefore, in section 3.3, we will explain the implementation of cloud computing to harness the potential of analyzing big data.

From the CC repository, we need to query the historical information by using the website address. The CC repository provides index files which contain Uniform Resource Locator (URL) and the file location. Web Archives (WARC) is the standardized format to archive the HTML of the web. For example, website companies such as www.shell.com had 3,200 web pages with unique URLs when it was crawled in July 2021. Hence, in one WARC record, we need to decompose the thousands of websites into different files to represent one file for one web page.

3.2 Methodology

3.2.1 Common Crawl Index

Common Crawl provides the index, which contains the URLs and other detail such as date, registered domain, and the file segment and offset. This index can be used to find the content based on the unique URLs. To describe the index file, an capture of the index file is depicted on Appendix 1. Common Crawl provides three types of datasets, i.e. WARC, WET, and WAT format. WARC format is the raw data from the crawl, which contains the whole website, including styling and metadata. WET files store the data, including the metadata and text. And WAT format contains only plain text.

The file segment consists of thousands of web archives, and the offset will help to locate the specific page. For this research, we only require plain text from each page. It gives the advantage where the file has the smallest size. Unfortunately, the index does not provide the offset for WAT format. The index only provides the offset from the WARC format. Therefore, we will crawl the data from the WARC files and only extract the text afterwards. This can be done during the crawling process by using the python library.

The registered domain on the web archive is used as the domain target because it should be exactly the same with the registered domain from ORBIS data. To join the table between the ORBIS data and CC Index, we utilize Amazon Athena service. During this process, we will obtain the new files containing the Index for the entire company website. Because we are going to build longitudinal data, we will crawl the data from several different dates. The number of URLs that has been targeted is summarized on Table 1.

For the start point, we choose based on the common news in which the corona or COVID-19 cases emerges in Wuhan, China. We expect that before January 2020, the COVID-19 case has not reached the common news. The first case of COVID-19 arrived in the EU on 24 January 2020. We choose January 2022 as the last date of data retrieval because we assume that the COVID-19 case has become "the new normal" in this sense.

Table 1. Number of Domains and URLs

No	Crawl Date	Domains (n)	URLs (n)
1	January 2020	172,291	120,629,130
2	May 2020	173,079	97,591,444
3	August 2020	174,043	103,215,772
4	October 2020	172,260	108,184,453
5	January 2021	173,949	118,628,811
6	April 2021	172,008	112,572,785
7	July 2021	177,046	108,417,562
8	October 2021	175,367	123,919,327
9	January 2022	173,419	113,523,891

The total number of URLs reach one billion web pages for the entire date. According to information in table 1, each date has a small change number of domains or approximately 2% from total domains. This indicates the number of domains tends to be stable over the period. The variety numbers of URLs might come from the dynamic condition of each website or unstable conditions while crawling the data. The availability of domains varies between 172,000 and 177,000, which represents the number of companies. The available domains are about 57.5-59.2% of the total number of companies provided by ORBIS data.

3.2.2 Data Queries Strategy

Before crawling, we need to observe the distribution of URLs among companies. We use the data from the July 2021 repository, which indicates the highest number of companies compared to data from another date. Although the number is not necessarily representing the exact number of companies, because of the dynamic condition during the crawl, we expect this number represent the companies for the crawling process.

Some companies have thousands of web pages, while other has less than a hundred pages. The top 10 domains with the most URLs are indicated in Table 2. Top 1,000 domains account for 87 million URLs. Half of total domains has less than 38 URLs. The histogram of distribution of URL is depicted in Figure 3. The horizontal axis is the number of URLs in the exponential class. The vertical axis is the frequency of companies which own the URLs.

Table 2. Top 10 Domain with highest URLs

No	Domain/Companies	Top URLs (n)
1	yahoo.com	1,372,835
2	google.com	927,933
3	elsevier.com	841,136
4	microsoft.com	581,680
5	columbia.edu	532,059
6	lefigaro.fr	467,453
7	apple.com	452,382
8	mpg.de	394,153
9	ku.dk	353,098
10	freepik.com	349,781

We want to make an efficient strategy which can reduce the number of crawled data without sacrificing the significant reduction. All crawled data must be correlated with the COVID-19 topic. We use the keywords on COVID-19 to find the relevant information. The list of keywords is provided in Appendix 2. The simplest way is we imagine how search engine such as Google or Yahoo works. We type a relevant keyword, and all relevant information will be shown. The search engines work by analyzing the URL, textual content, key content tags, and attributes (Gandour & Regolini, 2011). Furthermore, most of the companies we observed have their own website in which they present their policy measures against COVID-19. This type of website is the most relevant in our analysis

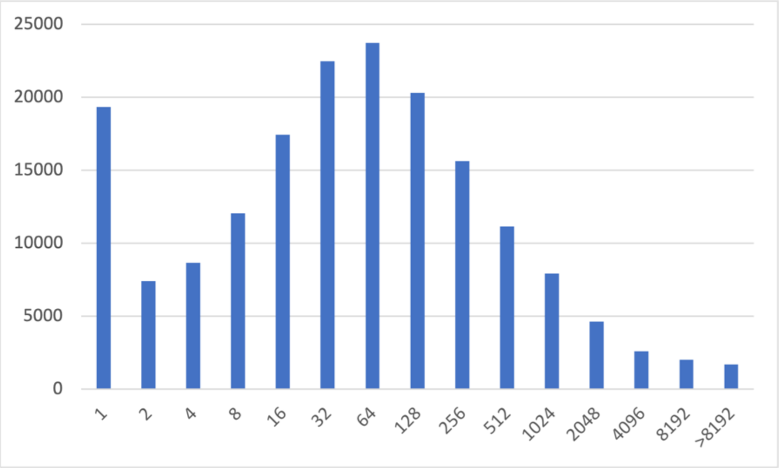


Figure 4. Distribution of URLs among Domains

In this case, the Common Crawl Index only provides the URLs, which also can be used by search engines to find the relevant information. Otherwise, we need to open the whole web page and search for the content that we are looking for. (Henzinger et al., 2000) considered URL sampling can be used to determine the properties of the entire website. Based on this research, we will use the keyword to filter the COVID-19 topic on the URL level. Opening the entire page one by one will increase the amount of related data, but it will take a lot of time and resources. The limitation of time scope of this research for searching from one billion pages is beyond the bounds of possibilities because of the bottleneck from the Common Crawl database.

3.3 Technology

3.3.1 Object Storage

Object storage is a storage to store data of many types, such as images, videos, or documents (Fazio et al., 2015). Easy way to imagine the object storage is similar to Google Drive provided by Google. We use Amazon Simple Storage Service (S3) to store all intermediate data and results in CSV, GZ, and Parquet format. Common Crawl is also stored on S3 through the Amazon Public Dataset program.

3.3.2 Serverless technology

Serverless technology is a cloud computing service which provides scalable computing without managing the server or instances (Chaudhary et al., 2017). Compared to common server, where we need to determine the specification such as core and ram size, serverless will count the optimum time with the scalable specifications. When the data processing needs high computing resources, the serverless will automatically distribute the process into many computers or servers.

We utilize two serverless services for this research, Amazon Athena and Amazon Glue. Athena is a serverless service to query data and analyzes big data in Amazon S3 by using the SQL programming language (Amazon, 2022a). Appendix 3 provides the SQL code during the URLs query process. Several previous studies implement PySpark during the index query process. Utilizing Athena cut the programming time and significantly faster the process. This process provides the whole web index from the registered domain.

After data has been crawled, we can visualize and edit it through AWS Glue Databrew. AWS Glue Databrew is designed as a data preparation tool to clean and normalize data for analyzing or machine learning (Amazon, 2022b). The advantage of Glue Databrew is we can create data connections to several sources such as S3 and RDS. Also, we can give the transformation and tasks to all the data without writing any code.

3.3.3 Map Reduce

Map reduce is a programming method and framework to deal with large datasets using a parallel or distributed algorithm (Dean & Ghemawat, 2008). The breakthrough of this method is we can run a process through multiple computers. When a personal computer runs all the process using the processor inside the personal computer, map-reduce allow us to run on multiple processors. By using a cloud computing service, the distributed process will be easy because we don't need to upscale the physical computer.

We used Amazon Elastic Map Reduce (EMR) during this research. Other than cloud computing service, the advantage of EMR is built with the open source big data frameworks such as Apache Spark and Apache Hive. We use the Apache Spark environment and Python programming language for Spark called PySpark. Without the built-in framework, the programmer needs to set all the environments, which will take more time before running the program. Appendix 4 provides the Pyspark code for crawling the data from the URL list on CSV to S3 with GZ format.

Implementation of map-reduce is over specification in this research. This arises because of the bottleneck during the crawling process, where the crawling process from the Common Crawl database is very slow. We know the bottleneck after we implement the process using Python Spark, which can also be done by using Python 2 and Python 3. The crawling process can be done by using normal instances without using a big data framework and will save cost from computational expenses. Compared to serverless, the EMR runs on a normal server when we pay, depending on the capacity/hour usage.

3.3.4 Database

Database is an organized of structured data which is stored on a computer system and controlled by a database management system (DBMS) (Oracle, 2022). We use AWS Relational Database

(RDS) as the DBMS in the cloud. We use a PostgreSQL database which can be operated by SQL programming language. We use the PGAdmin4 as Graphic User Interface (GUI) tool to manage the database. Appendix 5 provides the Pyspark code to convert the data from the GZ file and move it to the PostgreSQL database.

After we crawl the data using EMR, the data is saved as files on S3. This is a separate file and not modifiable. We build a table and combine all the files and the data through the database. The data on the DBMS can be connected to other datasets using AWS Glue Databrew to join attributes with other datasets.

4. Result

This chapter provides an overview of the COVID-19 keywords in companies' website texts. The study will measure the adaptation of COVID-19 on the companies across time and among sectors and countries. Section 4.1 explains the result on A-9 and A-13. Section 4.2 develops the potential of the information from the website to measure certain policies related to companies. In this case, we try to count the keyword "Work From Home" as the common policy during the outbreak.

4.1 Domain Analysis

4.1.1 COVID-19 Topic Presence on Website

Figure 5 illustrates the trend of URLs for each quartal from January 2020 until January 2022. Based on the query result, we find 4,283,604 contains COVID-19 keywords from 1,006,683,175 total URLs or 0.43% website page mentioning about COVID-19. Overall, the companies mention COVID-19 after Q1. During the pandemic period, several sectors such as business services, media & broadcasting, and printing are the top sectors which mention COVID-19 keywords on their website. Appendix 6 provide details about domain analysis data

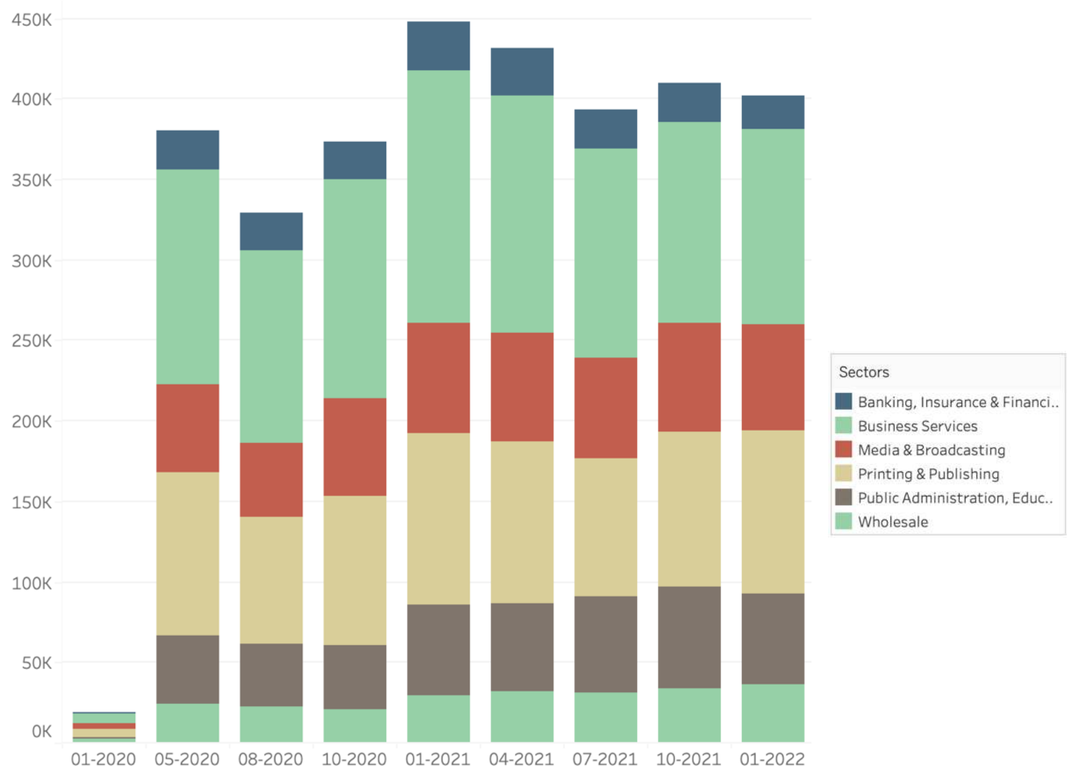
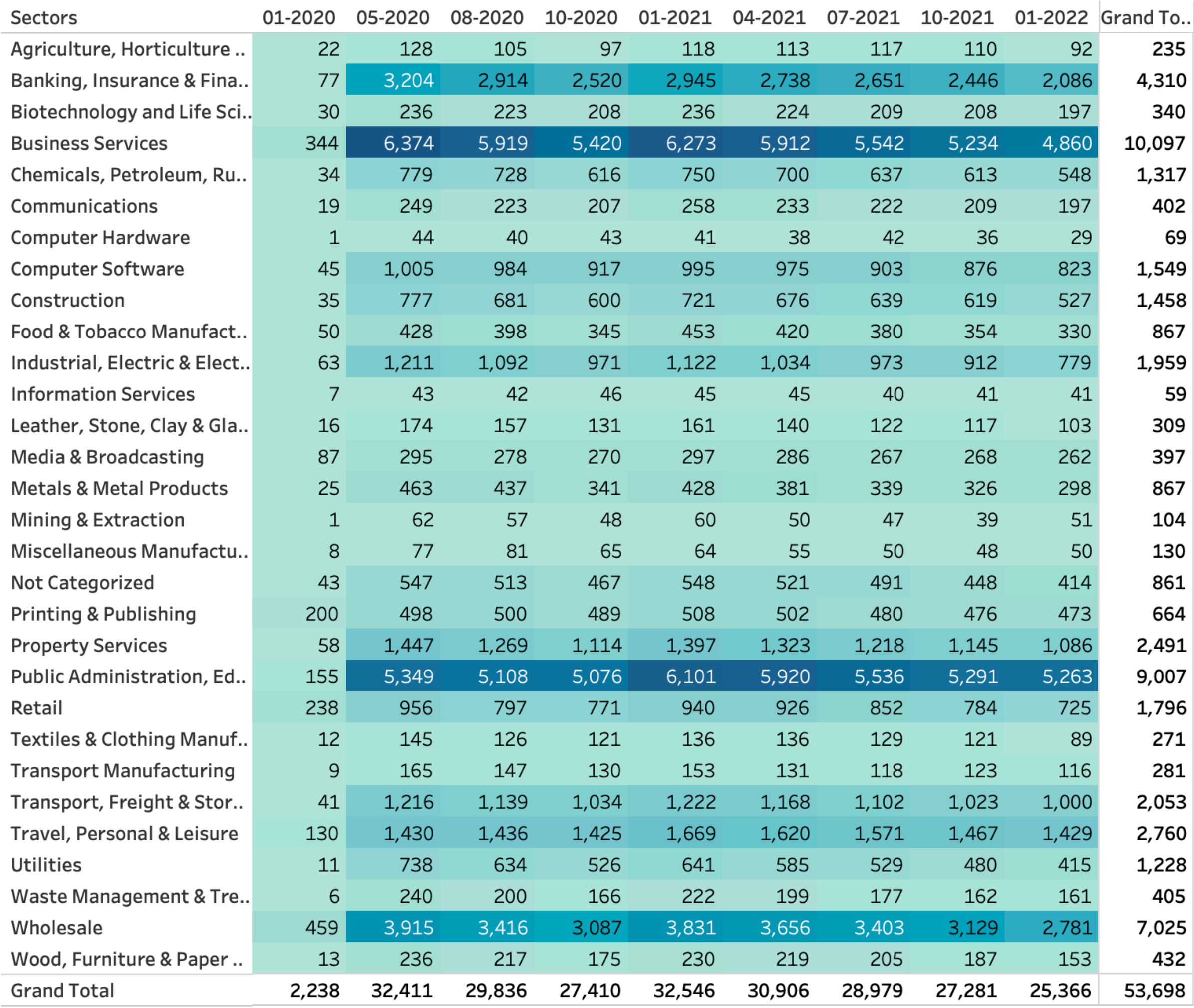


Figure 5 Number of URLs across sectors

4.1.2 COVID-19 Topic Presence Across Sectors

Figure 6 illustrates the proportion of companies that mention COVID-19 across sectors. 53,698 companies mention about COVID-19 via their website. We are able to capture around 30% of total companies. Business services, public administration, banking, and wholesale sectors contribute the most companies mention COVID-19. These sectors account for 49% of total industries. The high share of these sectors could be the early indication that companies in these sectors are most affected by COVID-19.

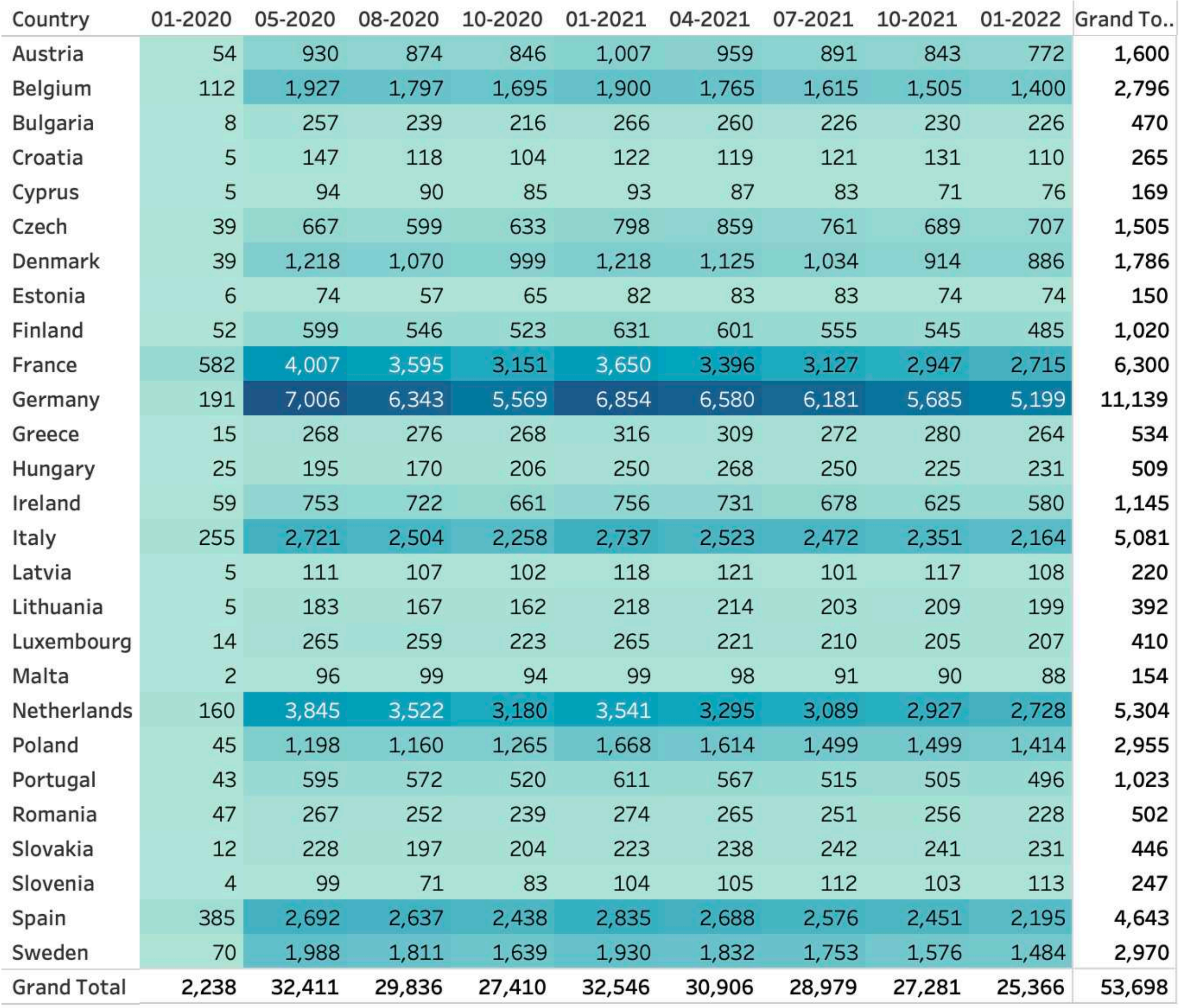


Sectors	01-2020	05-2020	08-2020	10-2020	01-2021	04-2021	07-2021	10-2021	01-2022	Grand To..
Agriculture, Horticulture ..	22	128	105	97	118	113	117	110	92	235
Banking, Insurance & Fina..	77	3,204	2,914	2,520	2,945	2,738	2,651	2,446	2,086	4,310
Biotechnology and Life Sci..	30	236	223	208	236	224	209	208	197	340
Business Services	344	6,374	5,919	5,420	6,273	5,912	5,542	5,234	4,860	10,097
Chemicals, Petroleum, Ru..	34	779	728	616	750	700	637	613	548	1,317
Communications	19	249	223	207	258	233	222	209	197	402
Computer Hardware	1	44	40	43	41	38	42	36	29	69
Computer Software	45	1,005	984	917	995	975	903	876	823	1,549
Construction	35	777	681	600	721	676	639	619	527	1,458
Food & Tobacco Manufact..	50	428	398	345	453	420	380	354	330	867
Industrial, Electric & Elect..	63	1,211	1,092	971	1,122	1,034	973	912	779	1,959
Information Services	7	43	42	46	45	45	40	41	41	59
Leather, Stone, Clay & Gla..	16	174	157	131	161	140	122	117	103	309
Media & Broadcasting	87	295	278	270	297	286	267	268	262	397
Metals & Metal Products	25	463	437	341	428	381	339	326	298	867
Mining & Extraction	1	62	57	48	60	50	47	39	51	104
Miscellaneous Manufactu..	8	77	81	65	64	55	50	48	50	130
Not Categorized	43	547	513	467	548	521	491	448	414	861
Printing & Publishing	200	498	500	489	508	502	480	476	473	664
Property Services	58	1,447	1,269	1,114	1,397	1,323	1,218	1,145	1,086	2,491
Public Administration, Ed..	155	5,349	5,108	5,076	6,101	5,920	5,536	5,291	5,263	9,007
Retail	238	956	797	771	940	926	852	784	725	1,796
Textiles & Clothing Manuf..	12	145	126	121	136	136	129	121	89	271
Transport Manufacturing	9	165	147	130	153	131	118	123	116	281
Transport, Freight & Stor..	41	1,216	1,139	1,034	1,222	1,168	1,102	1,023	1,000	2,053
Travel, Personal & Leisure	130	1,430	1,436	1,425	1,669	1,620	1,571	1,467	1,429	2,760
Utilities	11	738	634	526	641	585	529	480	415	1,228
Waste Management & Tre..	6	240	200	166	222	199	177	162	161	405
Wholesale	459	3,915	3,416	3,087	3,831	3,656	3,403	3,129	2,781	7,025
Wood, Furniture & Paper ..	13	236	217	175	230	219	205	187	153	432
Grand Total	2,238	32,411	29,836	27,410	32,546	30,906	28,979	27,281	25,366	53,698

Figure 6. Number of Companies mentioning COVID-19 across sector

4.1.3 COVID-19 Topic Presence Across Countries

The figure 7 illustrates the proportional of companies mention COVID-19 across countries. The companies from France, Germany, and Netherlands are the highest number of companies which mention COVID-19. The leap number of mentioning COVID-19 is indicated between the Q1-2020 and Q-2020. Companies awareness of COVID-19 pandemic can be obtained from website information. Companies mention the most COVID-19 keywords on Q2-2020 and Q1-2021. This can be related with other information such as the first wave and second wave of COVID-19.



Country	01-2020	05-2020	08-2020	10-2020	01-2021	04-2021	07-2021	10-2021	01-2022	Grand To..
Austria	54	930	874	846	1,007	959	891	843	772	1,600
Belgium	112	1,927	1,797	1,695	1,900	1,765	1,615	1,505	1,400	2,796
Bulgaria	8	257	239	216	266	260	226	230	226	470
Croatia	5	147	118	104	122	119	121	131	110	265
Cyprus	5	94	90	85	93	87	83	71	76	169
Czech	39	667	599	633	798	859	761	689	707	1,505
Denmark	39	1,218	1,070	999	1,218	1,125	1,034	914	886	1,786
Estonia	6	74	57	65	82	83	83	74	74	150
Finland	52	599	546	523	631	601	555	545	485	1,020
France	582	4,007	3,595	3,151	3,650	3,396	3,127	2,947	2,715	6,300
Germany	191	7,006	6,343	5,569	6,854	6,580	6,181	5,685	5,199	11,139
Greece	15	268	276	268	316	309	272	280	264	534
Hungary	25	195	170	206	250	268	250	225	231	509
Ireland	59	753	722	661	756	731	678	625	580	1,145
Italy	255	2,721	2,504	2,258	2,737	2,523	2,472	2,351	2,164	5,081
Latvia	5	111	107	102	118	121	101	117	108	220
Lithuania	5	183	167	162	218	214	203	209	199	392
Luxembourg	14	265	259	223	265	221	210	205	207	410
Malta	2	96	99	94	99	98	91	90	88	154
Netherlands	160	3,845	3,522	3,180	3,541	3,295	3,089	2,927	2,728	5,304
Poland	45	1,198	1,160	1,265	1,668	1,614	1,499	1,499	1,414	2,955
Portugal	43	595	572	520	611	567	515	505	496	1,023
Romania	47	267	252	239	274	265	251	256	228	502
Slovakia	12	228	197	204	223	238	242	241	231	446
Slovenia	4	99	71	83	104	105	112	103	113	247
Spain	385	2,692	2,637	2,438	2,835	2,688	2,576	2,451	2,195	4,643
Sweden	70	1,988	1,811	1,639	1,930	1,832	1,753	1,576	1,484	2,970
Grand Total	2,238	32,411	29,836	27,410	32,546	30,906	28,979	27,281	25,366	53,698

Figure 7. Number of Companies mentioning COVID-19 across nation

4.2 Content Analysis

The figure illustrates the count of "work from home" in the companies' websites across sectors. The term "work from home" is strongly related to the company policy to adapt to the COVID-19. Despite many company policies to adapt to the pandemic, we assess that the term "work from home" would reflect most of the companies' working adaptations. The second quarter of 2020 has the highest count of "work from home". We stipulate that this is due to the rapidly increasing COVID-19 cases in the European countries. At that time, the companies must abide by the lockdown rule.

We also represent the count of the term "work from home" in different sectors. Business services sectors dominates in most of the time. This might be because those companies in the services sector tend to be able to adjust their business operations into the "work from home". On the contrary, other sectors tend to be more difficult to adjust their working process into the "work from home". The limitation on this study is several website using different languages. Therefore, we need to divide into many different language to obtain the full information

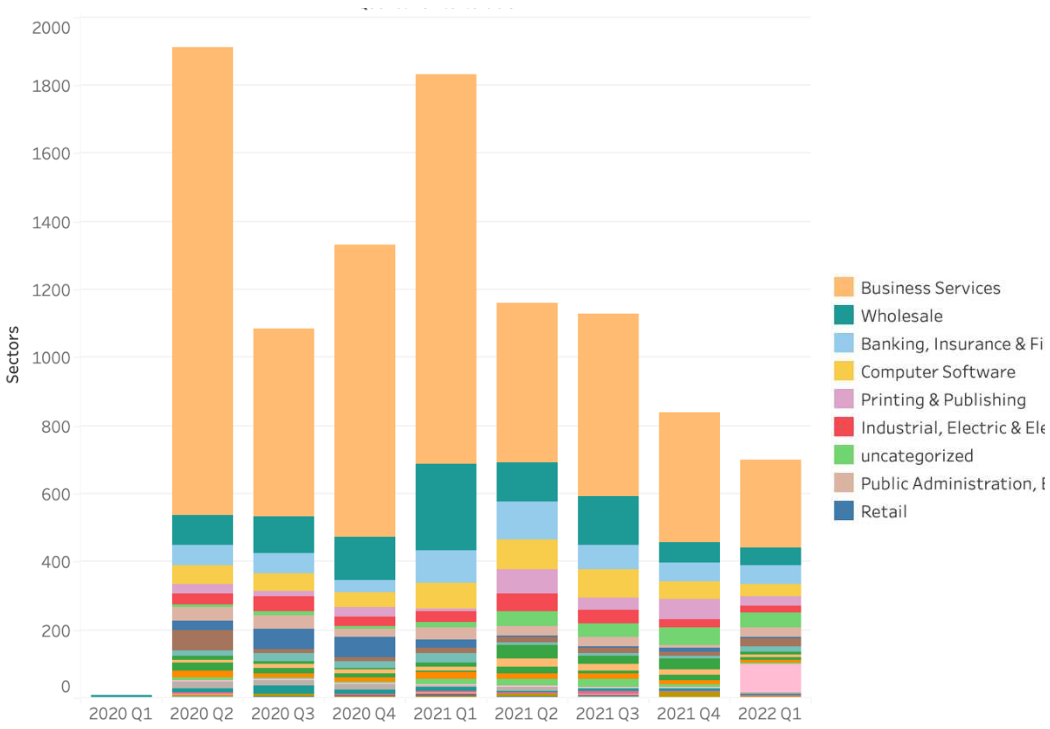


Figure 8. Number of Sector mentioning "work from home"

4.3 Accessing the data

The raw data is the main part of the deliverable on this study. To deliver the metadata via CSV is not recommended because the text contain more than a CSV cell can handle. The data is stored on PostgreSQL on AWS Relational Database. To access the data, we require a GUI software and we recommend PGAdmin. Below is the address to connect to database.

- Hostname: database-crawl-13.cxvwrotyzkne.us-east-1.rds.amazonaws.com
- Port: 5432
- Username: postgres
- Password: postgres

5. Conclusion and Limitation

5.1 Conclusion

This study unlock the new potential of revealed preference data sources. Specifically, we obtain the information about the COVID-19 through companies website. Different from previous study, this research provides the novelty of data crawling process with specific keywords and specific entity. To produce this data, the big data framework is required during the process. Serverless and Map Reduce are introduced as the latest technology to run big data project. We implement cloud computing service to provide seamless flow.

Moreover, this study can build the panel data with quartal temporal resolution from the beginning of pandemic until the new normal condition. We capture that 0.49% from total website pages mention about COVID-19. Also, we capture around 53,698 in European Union or 30% of total companies from ORBIS data which mention COVID-19 keywords. Business services, public administration, and banking accounts for 49% from total companies and become the top sector which mention COVID-19. Companies from French, Germany, and Netherlands also become the top nation which mention COVID-19.

Also, we try to analyse the content using the word "work from home". This can be an example of the measurement of company policies that may be issued to deal with COVID-19. Further study can be conducted with this data to find the causality of certain topic which is explained by the information from the website.

5.2 Limitation and Further Studies

Several limitation appears during the process of this study. We divide the obstacle as methodology limitation and technical limitation. On the methodology limitation, we do not identify deeper the type of information which appears on the company website. For advertising sector, the information probably represents other entities such as news or relay information and does not represent their own information. For health sectors, the website might represent the recommendation or the situation of spread of COVID-19 disease. Further investigation is required to find context of the information with our purpose.

Furthermore, we do not implement the causal inference research where we can explain the phenomena based on the data that has been crawled. This research is more focus on how we crawl the data. Future study can use this dataset to find the causality between the COVID-19 information on website to other variable. Some interesting topic can be proposed such as explain the company policy during COVID-19 or hiring and layoff process during the spread of COVID-19 pandemic.

For the technical limitation, we implement many versions of open source programming and technology. The challenge comes from the different version of environment based on the previous research or project with the current condition. It takes more time to do some research for integrating or comply with the system. We found some problem arises because we cannot find the compatible version during integration process.

Moreover, the solution when trouble or bug occurs during the implementation cannot be found on formal reference such as books or papers. We rely on the forum and QnA in several websites like Google Groups or Stack Overflow. The solution sometimes not reproduceable and outdated with the current technology version. Problem solving requires a combination of all available partial information.

For the further study, our tech stack can be implemented for similar research and minimize the technical trial and error. The tech stack using a mature cloud computing service and several process does not require a programming. Also, the technology is widely used on the developer community and some experts or peers can solve the question on the open forum.

I. Reference

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining Text Data*. Springer US. https://doi.org/10.1007/978-1-4614-3223-4

Amazon. (2022a). *Amazon Athena*. https://aws.amazon.com/athena/

Amazon. (2022b). *AWS Glue Databrew*. https://aws.amazon.com/glue/features/databrew/

Baldwin, R. E., Weder, B., & Centre for Economic Policy Research (Great Britain). (2020). *Economics in the time of COVID-19*. CEPR Press.

Benveniste, L. M., Ljungqvist, A., Wilhelm, W. J., & Yu, X. (2003). Evidence of Information Spillovers in the Production of Investment Banking Services. *The Journal of Finance*, *58*(2), 577–608. https://doi.org/10.1111/1540-6261.00538

Buchheim, L., Dovern, J., Krolage, C., & Link, S. (2022). Sentiment and firm behavior during the COVID-19 pandemic. *Journal of Economic Behavior & Organization*, *195*, 186–198. https://doi.org/10.1016/j.jebo.2022.01.011

Bulao, J. (2022). *How Much Data Is Created Every Day in 2022*. https://techjury.net/blog/how-much-data-is-created-every-day/#gref

Chaudhary, S., Somani, G., & Buyya, R. (Eds.). (2017). *Research Advances in Cloud Computing*. Springer Singapore. https://doi.org/10.1007/978-981-10-5026-8

Chetty, R., Friedman, J., Hendren, N., Stepner, M., & Team, T. O. I. (2020). *The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data* (No. w27431; p. w27431). National Bureau of Economic Research. https://doi.org/10.3386/w27431

CommonCrawl.org. (2022). *Https://commoncrawl.org/big-picture/what-we-do/*. https://commoncrawl.org/big-picture/what-we-do/

Constantinides, E., & Fountain, S. J. (2008). Web 2.0: Conceptual foundations and marketing issues. *Journal of Direct, Data and Digital Marketing Practice*, *9*(3), 231–244. https://doi.org/10.1057/palgrave.dddmp.4350098

Daas, P. J. H., & van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, *36*(4), 1239–1251. https://doi.org/10.3233/SJI-200627

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113. https://doi.org/10.1145/1327452.1327492

Deloitte. (2020). *COVID-19—Managing supply chain risk and disruption*. 20.

Dillon, Wu, C., & Chang, E. (2010, April). Cloud Computing: Issues and Challenges. *2010 24th IEEE International Conference on Advanced Information Networking and Applications*.

El-Showk, S. (2018). *A growing proportion of global culture exists only online, presenting a challenge for those tasked with maintaining the historical record.* 3.

Fazio, M., Celesti, A., Puliafito, A., & Villari, M. (2015). Big Data Storage in the Cloud for Smart Environment Monitoring. *Procedia Computer Science*, *52*, 500–506. https://doi.org/10.1016/j.procs.2015.05.023

Gamerschlag, R., Möller, K., & Verbeeten, F. (2011). Determinants of voluntary CSR disclosure: Empirical evidence from Germany. *Review of Managerial Science*, *5*(2–3), 233–262. https://doi.org/10.1007/s11846-010-0052-3

Gandour, A., & Regolini, A. (2011). Web site search engine optimization: A case study of Fragfornet. *Library Hi Tech News*, *28*(6), 6–13. https://doi.org/10.1108/07419051111173874

Goos, G., Hartmanis, J., van Leeuwen, J., Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Kobsa, A., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Rangan, C. P., & Steffen, B. (2009). *Lecture Notes in Computer Science*. 726.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & de Jong, F. (2011). Polarity analysis of texts using discourse structure. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11*, 1061. https://doi.org/10.1145/2063576.2063730

Heinze, N., & Hu, Q. (2006). The evolution of corporate web presence: A longitudinal study of large American companies. *International Journal of Information Management*, *26*(4), 313–325. https://doi.org/10.1016/j.ijinfomgt.2006.03.008

Henzinger, M. R., Heydon, A., Mitzenmacher, M., & Najork, M. (2000). On near-uniform URL sampling. *Computer Networks*, *33*(1–6), 295–308. https://doi.org/10.1016/S1389-1286(00)00055-4

Internetlivestat. (2022). https://www.internetlivestats.com/total-number-of-websites/

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*, 995–1004. https://doi.org/10.1109/HICSS.2013.645

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

Krüger, M., Kinne, J., Lenz, D., & Resch, B. (n.d.). *The Digital Layer: How Innovative Firms Relate on the Web*. 14.

Lawinsider. (2022). https://www.lawinsider.com/dictionary/company-websites#

Loayza, N. V., & Pennings, S. (2020). *Macroeconomic Policy in the Time of COVID-19: A Primer for Developing Countries*. 9.

Maital, S., & Barzani, E. (2020). *Global Economic Impact*. 12.

Mason, E. S. (1939). *Price and Production Policies of Large-Scale Enterprise*. 15.

Michael, K., Parnian, S., Jakob, R., & Martin, W. (2022). *How were Companies Affected During the First and Second Waves of the Corona Pandemic in Switzerland?: An Analysis based on KOF Survey Data, Short Term Work and Company Websites* (p. 70 p.) [Application/pdf]. ETH Zurich. https://doi.org/10.3929/ETHZ-B-000527411

Nicholson, W., & Snyder, C. (2008). *Microeconomic theory: Basic principles and extensions* (10th ed). Thomson Business and Economics.

Niu, J. (2012). *Functionalities of Web Archives*. 16.

Oracle. (2022). *What is Database?* https://www.oracle.com/database/what-is-database/

Robbins, S. S., & Stylianou, A. C. (2003). Global corporate web sites: An empirical investigation of content and design. *Information & Management*, *40*(3), 205–212. https://doi.org/10.1016/S0378-7206(02)00002-2

Shen, H., Fu, M., Pan, H., Yu, Z., & Chen, Y. (2020). The Impact of the COVID-19 Pandemic on Firm Performance. *Emerging Markets Finance and Trade*, *56*(10), 2213–2230. https://doi.org/10.1080/1540496X.2020.1785863

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., & Zhou, M. (2016). Sentiment Embeddings with Applications to Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(2), 496–509. https://doi.org/10.1109/TKDE.2015.2489653

Trauer, J., Pfingstl, S., Finsterer, M., & Zimmermann, M. (2021). Improving Production Efficiency with a Digital Twin Based on Anomaly Detection. *Sustainability*, *13*(18), 10155. https://doi.org/10.3390/su131810155

Worlodometer. (2022, January 31). https://www.worldometers.info/coronavirus/

II. Appendix

Appendix 1. Common Crawl Index Structure (with sample)

url	https://www.cgi.com/maroc/fr/article/carrieres/cov d-19-coronavirus-mise-jour-pour-les-candidats
url_host_name	www.cgi.com
url_host_registered_domain	cgi.com
warc_filename	crawl-data/CC-MAIN-2021-31/segments/1627046153966.52/warc/CC-MAIN-20210730091645-20210730121645-00072.warc.gz
warc_record_offset	710692051
warc_record_end	710745551
crawl_date	CC-MAIN-2021-31
subset	warc

The warc files are divided into 100 segments for each date. Each segment contain the WARC files as a text with millions random website. To crawl the specific URL on the sample, we use the warc_record_offset and warc_record_end to only get the data between the character 710,692,051st until character 710,745,551st. The sample URL has 5.350 characters inside the WARC file.

Appendix 2. COVID-19 Related Keywords

- Universal Keywords

Keywords	covid, covid 19, covid-19, corona, coronavirus, sars cov2, delta, omicron, pandemic

- Specific Related COVID-19 Keywords based on language in EU

Language	**Specific unique keywords**
Bulgarian	Корона, Пандемия,
Croatian	pandemija
Czech	Korona, Pandemický
Danish	pandemi
Finland	pandeeminen
France	Couronne, pandémie
Greece	Πανδημία, κορωνοϊός
Hungary	világjárvány
Ireland	Coróin, paindéim
Italy	pandemia
Latvia	pandēmija
Dutch	pandemie
Portugal	coroa
Slovak	koróna
Sweden	pandemisk

Appendix 3. SQL Code for query the Common Crawl Index

```
-- request 1, create a database
CREATE EXTERNAL TABLE IF NOT EXISTS ccindex (
  url_surtkey                   STRING,
  url                           STRING,
  url_host_name                 STRING,
  url_host_tld                  STRING,
  url_host_2nd_last_part        STRING,
  url_host_3rd_last_part        STRING,
  url_host_4th_last_part        STRING,
  url_host_5th_last_part        STRING,
  url_host_registry_suffix      STRING,
  url_host_registered_domain    STRING,
  url_host_private_suffix       STRING,
  url_host_private_domain       STRING,
  url_protocol                  STRING,
  url_port                      INT,
  url_path                      STRING,
  url_query                     STRING,
  fetch_time                    TIMESTAMP,
  fetch_status                  SMALLINT,
  content_digest                STRING,
  content_mime_type             STRING,
    content_mime_detected         STRING,
  content_charset               STRING,
  content_languages             STRING,
  warc_filename                 STRING,
  warc_record_offset            INT,
  warc_record_length            INT,
  warc_segment                  STRING)
PARTITIONED BY (
  crawl                         STRING,
  subset                        STRING)
STORED AS parquet
LOCATION 's3://commoncrawl/cc-index/table/cc-main/warc/';


--request 2, repair database
MSCK REPAIR TABLE ccindex

--request 3, create table from list of companies in csv
CREATE EXTERNAL TABLE IF NOT EXISTS ccindex.scrape500(
  `row_no` string,
  `companyname` string,
  `websiteaddress` string,
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ";"
LINES TERMINATED BY "\n"
LOCATION 's3://dirga/'
```

```
TBLPROPERTIES (
'skip.header.line.count' = '1'
);

--request 4, joining the csv with the warcfile index from commoncrawl
Create table websiteall_crawl_2022_05 AS
SELECT url,
       url_host_name,
       url_host_registered_domain,
       warc_filename,
       warc_record_offset,
       warc_record_offset + warc_record_length as warc_record_end,
       crawl,
       subset
FROM ccindex.ccindex
JOIN ccindex.websiteall ON ccindex.ccindex.url_host_registered_domain =
ccindex.websiteall.websiteaddress
WHERE crawl = 'CC-MAIN-2022-05'
  AND subset = 'warc'
```

This SQL code is a development from https://commoncrawl.org/2018/03/index-to-warc-files-and-urls-in-columnar-format/

Appendix 4. Code for access the Common Crawl Data with specific segment and record offset and length

```
import boto3
import requests
import csv
# Import library
```

```
s3 = boto3.client(
   's3',
   aws_access_key_id='ACCESS_KEY_ID',
   aws_secret_access_key=’SECRET_ACCESS_KEY’
)
# access s3 using secret access key
```

```
def list_csv_in_folder(bucket_name, folder_name):
   file_names = []
   response = s3.list_objects_v2(Bucket=bucket_name, Prefix=folder_name)
   files = response.get("Contents")
   for file in files:
      if file['Size'] > 0 and '.csv' in file['Key']:
         file_names.append(file['Key'])
   return file_names
# define list csv (data contains warc filename and offset)


def list_gz_in_folder(bucket_name, folder_name):
   file_names = []
   response = s3.list_objects_v2(Bucket=bucket_name, Prefix=folder_name)
   files = response.get("Contents")
   for file in files:
      if file['Size'] > 0 and '.gz' in file['Key']:
         file_names.append(file['Key'])
   return file_names
# define list gz in S3 after data has been crawled
```

```
def url_to_gz(url, headers, folder_name, offset):
   local_filename = url.split('/')[-1]
   with requests.get(url, headers=headers, stream=True) as r:
      s3.upload_fileobj(r.raw, "url-contain-keywords",
                  folder_name+'/'+offset+local_filename  )
   return folder_name+'/'+ offset+local_filename


def csv_to_gz(bucket_name, file_name, output_folder):
   obj = s3.get_object(Bucket=bucket_name, Key=file_name)
```

```
    f = obj['Body'].read().decode('utf-8').splitlines()
    reader = csv.DictReader(f)
    line_count = 0
    keys = []
    for row in reader:
        if line_count > 0:  # not header
            s3_key = url_to_gz('https://data.commoncrawl.org/'+row['warc_filename'],
                        {"range": f"bytes= {row['warc_record_offset']}-
{row['warc_record_end']}"},
                        output_folder, row['warc_record_offset'])
            keys.append(s3_key)
        line_count += 1
    return keys
```

```
for file_name in list_csv_in_folder("url-contain-keywords", 's3_key'):
    csv_to_gz('url-contain-keywords', file_name, ‘s3_key’)

#file input and output should be on the same Bucket
```

Appendix 5. Code for convert the GZ files into database

```
# Importing necessary libraries
import boto3
import requests
import csv

from io import BytesIO
from warcio.archiveiterator import ArchiveIterator
from bs4.dammit import EncodingDetector
from bs4 import BeautifulSoup
```

```
# Initializing s3 access
s3 = boto3.client(
    's3',
    aws_access_key_id='[**************]',
    aws_secret_access_key=’****************’
)
```

```
# Defining functions for removing blank lines and get text from html

def remove_blank_line(article):
    lines = article.split("\n")
    non_empty_lines = [line for line in lines if line.strip() != ""]

    string_without_empty_lines = ""
    for line in non_empty_lines:
        string_without_empty_lines += line + "\n"

    return string_without_empty_lines


def gz_to_content(bucket_name, s3_key):
    # rangereq = f'bytes={offset}-{end}' # still necessary ?
    # response = s3.get_object(Bucket='commoncrawl',Key=s3_key,Range=rangereq)
    response = s3.get_object(Bucket=bucket_name, Key=s3_key)
    record_stream = BytesIO(response["Body"].read())
    record = ArchiveIterator(record_stream)
    for record in ArchiveIterator(record_stream):
        content = record.content_stream().read()
        encoding = EncodingDetector.find_declared_encoding(
            content, is_html=True)
        soup =  BeautifulSoup(content, "html.parser", from_encoding=encoding)
        return remove_blank_line(soup.get_text())
```

```
# Initializing connections to database
```

```
import psycopg2

conn = psycopg2.connect(
   host="HOST",
   database="DATABASE",
   user="USERNAME",
   password="*****")
```

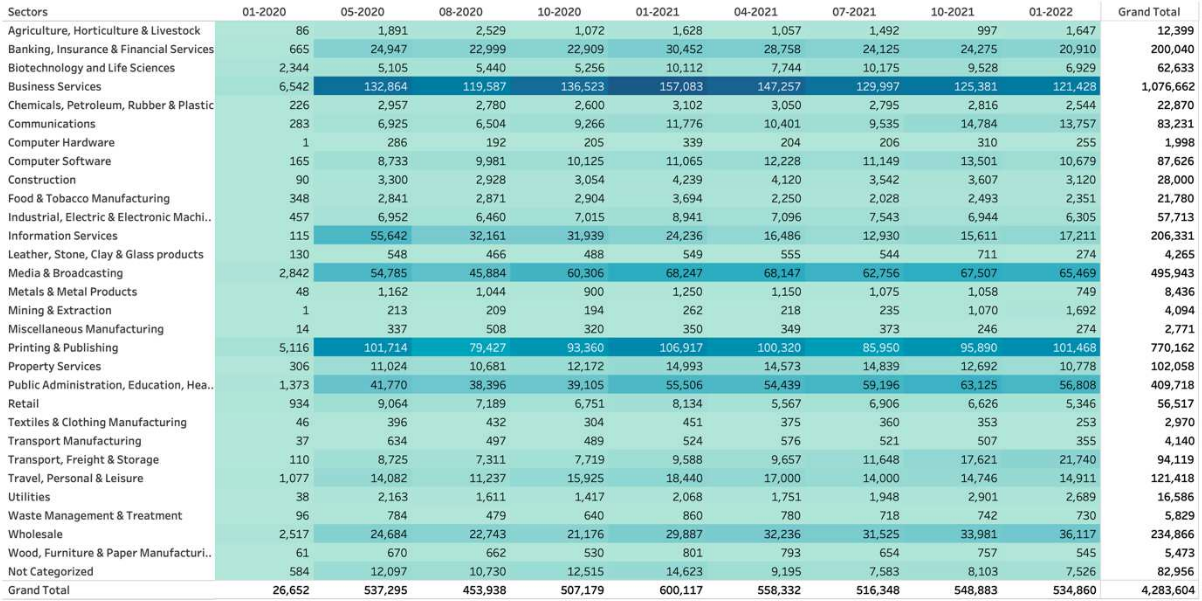
```
# Defining functions for each GZ file to generate the text/content
def content_to_row( pkey, mainkey,  crawlcontent):
   cur = conn.cursor()
   cur.execute(
      "INSERT INTO TABLE ( pkey, mainkey,  crawlcontent) VALUES (%s, %s, %s)", (pkey,
mainkey,     crawlcontent))
   conn.commit()
   cur.close()
   print('Running')
#TABLE is the name of the table in the database
```

```
# Paginating the s3 bucket since the files is more than one thousands
paginator = s3.get_paginator('list_objects_v2')
pages = paginator.paginate(Bucket='url-contain-keywords', Prefix='urlfilter-crawl-2020-
34_24Jul2022_1658699213012-output')
```

```
# Listing the gz in the s3 folder
listgz = []
for page in pages:
   pagee = page['Contents']
   for pag in pagee:
      listgz.append(pag['Key'])
```

```
# Executing the process of generating "content" into the database
pkey = 0
for gz in listgz:
      try:
         content = gz_to_content('url-contain-keywords', gz)
         pkey +=1
         gz = gz.split('/')[-1]
         content_to_row(pkey, gz, content)
      except:
         continue
```

Appendix 6. Number of URLs across sector



Sectors	01-2020	05-2020	08-2020	10-2020	01-2021	04-2021	07-2021	10-2021	01-2022	Grand Total
Agriculture, Horticulture & Livestock	86	1,891	2,529	1,072	1,628	1,057	1,492	997	1,647	12,399
Banking, Insurance & Financial Services	665	24,947	22,999	22,909	30,452	28,758	24,125	24,275	20,910	200,040
Biotechnology and Life Sciences	2,344	5,105	5,440	5,256	10,112	7,744	10,175	9,528	6,929	62,633
Business Services	6,542	132,864	119,587	136,523	157,083	147,257	129,997	125,381	121,428	1,076,662
Chemicals, Petroleum, Rubber & Plastic	226	2,957	2,780	2,600	3,102	3,050	2,795	2,816	2,544	22,870
Communications	283	6,925	6,504	9,266	11,776	10,401	9,535	14,784	13,757	83,231
Computer Hardware	1	286	192	205	339	204	206	310	255	1,998
Computer Software	165	8,733	9,981	10,125	11,065	12,228	11,149	13,501	10,679	87,626
Construction	90	3,300	2,928	3,054	4,239	4,120	3,542	3,607	3,120	28,000
Food & Tobacco Manufacturing	348	2,841	2,871	2,904	3,694	2,250	2,028	2,493	2,351	21,780
Industrial, Electric & Electronic Machi..	457	6,952	6,460	7,015	8,941	7,096	7,543	6,944	6,305	57,713
Information Services	115	55,642	32,161	31,939	24,236	16,486	12,930	15,611	17,211	206,331
Leather, Stone, Clay & Glass products	130	548	466	488	549	555	544	711	274	4,265
Media & Broadcasting	2,842	54,785	45,884	60,306	68,247	68,147	62,756	67,507	65,469	495,943
Metals & Metal Products	48	1,162	1,044	900	1,250	1,150	1,075	1,058	749	8,436
Mining & Extraction	1	213	209	194	262	218	235	1,070	1,692	4,094
Miscellaneous Manufacturing	14	337	508	320	350	349	373	246	274	2,771
Printing & Publishing	5,116	101,714	79,427	93,360	106,917	100,320	85,950	95,890	101,468	770,162
Property Services	306	11,024	10,681	12,172	14,993	14,573	14,839	12,692	10,778	102,058
Public Administration, Education, Hea..	1,373	41,770	38,396	39,105	55,506	54,439	59,196	63,125	56,808	409,718
Retail	934	9,064	7,189	6,751	8,134	5,567	6,906	6,626	5,346	56,517
Textiles & Clothing Manufacturing	46	396	432	304	451	375	360	353	253	2,970
Transport Manufacturing	37	634	497	489	524	576	521	507	355	4,140
Transport, Freight & Storage	110	8,725	7,311	7,719	9,588	9,657	11,648	17,621	21,740	94,119
Travel, Personal & Leisure	1,077	14,082	11,237	15,925	18,440	17,000	14,000	14,746	14,911	121,418
Utilities	38	2,163	1,611	1,417	2,068	1,751	1,948	2,901	2,689	16,586
Waste Management & Treatment	96	784	479	640	860	780	718	742	730	5,829
Wholesale	2,517	24,684	22,743	21,176	29,887	32,236	31,525	33,981	36,117	234,866
Wood, Furniture & Paper Manufacturi..	61	670	662	530	801	793	654	757	545	5,473
Not Categorized	584	12,097	10,730	12,515	14,623	9,195	7,583	8,103	7,526	82,956
Grand Total	26,652	537,295	453,938	507,179	600,117	558,332	516,348	548,883	534,860	4,283,604