Travel and Economic growth Analysis

Kishore Lakshmanan
x20253583
Data Intensive Architectures
MSc in Data Analytics
National College of Ireland

***Abstract*—In the modern world, emergence of data is rapidly increasing in all the fields which tends to accumulate as a large data-set. Several business and scientific applications require the efficient processing of terabytes of data on a regular basis. With the allocated time, none of the existing database operations and tools/software were useful for analysis. There are several processing of data involved before turning into an proper result. One of the challenging thing is to check how quick the data is being retrieved and stored to get meaningful insights. So, everyone is focusing to gain knowledge in big data and their frameworks to process the data in an efficient way. To handle the large data efficiently, we introduce hadoop map-reduce concept to make fault-tolerant storage(hdfs) with high throughput. In this paper we summarizes how the MapReduce programming framework and Hadoop platform can be used to process massive amount of data. We apply these techniques on large data sets from information and communication sector(ICT) and passenger movement to see if there is a correlation between the economic growth of ICT sector and the Ireland's travel report.**

***Index Terms*—MapReduce, fault-tolerant, throughput, programming, information and communication sector, passenger movement, economic growth**

I. Introduction

Many people around the world created the huge volume of data. This datasets are populated in means of different formats, but the majority of it is unstructured one. The massive data is kept on hand to perform day-to-day analysis from the stored location. Furthermore, there are many data sources are freely available because of the machine populated data as newly emerged technology. This overcomes the obstacle faced by many humans to get proper datasets. These kind of Big Data were successeed in the market based on the resulted output [1]. The MapReduce Hadoop implementation has evolved significantly since its inception and become for significantly storing, organizing, and processing enormous amounts of data Hadoop MapReduce has emerged as a popular solution. [2] Tool for analyzing Big Data that is extremely effective and efficient. MapReduce is also great for digging the historical data and interpret the data.

The main aim of this project is to perform MapReduce method to analyze a large datasets and to conduct some correlation and post-processing visualization with the MapReduce results. Using this techniques we can showcase how the people movement from one country to another based on the ICT sector's value from different attributes.

II. Data

In this project we are using 2 datasets to do our analyzes. Both the datasets having more than 1M records and they are related in some ways to achieve the data intensive part. The following figure 1 shows the meta info about the datasets using tabular format.

Metadata	Datasets	
	ICT Data	Passenger Movement
Name of the Dataset	PREDICT 2021	CTM01-Passenger Movement
Repository	Europe	Ireland
Source	https://data.europa.eu/	https://data.cso.ie/
Publisher	Joint Research Centre	Data provided by Airport
Licence	Creative Commons Attribution 4.0 International	Central Statistics Office, Ireland
Access Rights	Public	Public
Format	Structured(CSV file)	Structured(CSV file)
Created Date	29-10-2021	NA
Modified Date	29-10-2021	30-09-2021

Fig. 1. Description of two data sets

The below content explains the detailed description of these 2 datasets.

ICT DATASET - This dataset is belongs to the European government and hosted on "data.europa.eu". Information and communication sector's value has been described in this dataset from the year 1995 to 2020. The licensing information of the data is provided in the above table and it has been published by Joint Research Centre. This is retrieved as a comma separated file and received as a publicly accessible. This dataset contains the following attributes.

- Variable Code (includes Price, Emp etc)
- Unit (includes National currency, 1000 person employed etc)
- Country Code
- Year
- Country
- Value
- Dataset type
- Classification (includes NACE Rev.2)
- Definition (includes Different sectors type)
- Sector code

- Description (includes Exchange Rate, Manufacturing, etc)

Out of these 11 attributes, we are going to use 4 in total to perform the task which are variable code,country, year and value.

PASSENGER MOVEMENT - The Irish government owns this dataset, which is hosted on "data.cso.ie." This report offers information on passenger travel in and out of Ireland from 2006 to 2020 as a month-by-month report. The Central Statistics Office of Ireland has licensed and published this. This is retrieved as a comma-separated file that is made publicly available. The following attributes are present in this dataset.

- Statistic
- Month
- Value
- Direction
- Foreign Airport
- Irish Airport
- Unit

With the above columns, 3 of them are interrelated to another dataset. They are Month which will be converted to year, value is the population count and the direction whether it is inward, outward or both.

A. Objectiveness of the research

This project has a many impact especially on the government revenue. It relates both the industry such as tourism and the information technology sector to find a proper conclusion on the economic crisis. Based on their value we can estimate how the growth would be in future days were calculated. Through the mapreduce concept we can find whether the incoming people will contribute to the economic growth or not and also based on the how much value they are spending to the employees makes sense to the more employment of the country which welcomes more people from outside to get benefited.

B. Use cases

The following items are the insights which will be used to analyse in this project.

1) Predict how will be the European economic growth flows based on the people movement.
2) Employee wages vs incoming people to the countries.
3) Based on gross value added of the sector, how much the country losses his own people.
4) In which the year, sector value(economic growth) and the population movement is more.

III. Methodology

The important part of our project move should be addressing the data quality after selecting the proper data set. Then the qualified data will be moved on to the MapReduce framework and resulted output should be used for finding correlation and visualization between them.

A. Stages Involved in Data Quality Validation

The purpose of the data quality assurance is to ensure the correctness, fulfill, proper checking of data, as well as to ensure that data is reliable, accurate, and meaningful. The major part of the project will be spend on discovered the data cleansing and preparation. [3]. The validation method for huge data is discussed in this section. The important processes in data validation are collecting, cleansing, transform, loading, and output reporting from the data.

The below steps and figure 2 illustrates the detailed analysis of data quality made in our project.

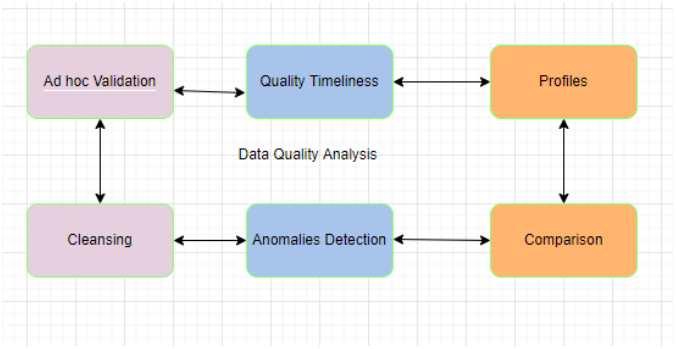


Fig. 2. Data Quality Analysis

1) Data Collection: It this stage, data is gathered from a variety of sources, and warehouses, databases, CSV files, excel spreadsheets, and file systems. In this project we have selected Comma separated files to run our analysis.
2) Loading of Data: Data is pushed into a big data platform, such as HDFS or a database, during this stage. Updates to extracted data are usually performed on a daily, weekly, or monthly basis, depending on the requirements. Here we are using Hadoop Distributed File-system for storing the large dataset for analysis.
3) Data Validation: The data quality is evaluated at this stage, including the datatype, numerical value, signed value, greater/lesser symbol and so on. In this project we have constrains in data types and conversion of month wise data to year etc.
4) Data Cleansing: The data is cleaned at this step by repairing or deleting corrupted or erroneous rows from a dataset / database. The objective is to find the data portions that are missing, erroneous, inaccurate, or irrelevant in data sets. Data cleansing is the process which takes more time for doing certain task in analytical and Big Data testing. In this project we have null records, hyphen coordinated rows exists. So we have to remove those to get better performance. This has been done through manually or by java code. In our case, we did it through java code by adding proper conditions in the relevant field.
5) Data Transformation: This stage involves converting a stack of data from one format to another format as like as source to destination.

6) Data Analysis and Reporting: The final stage of writing report is to display data validation result. With the purpose of obtaining usable information, data consistency, completeness, and other data quality aspects are thoroughly examined.In this we are going to use IBM spss and R to visualize the data to give proper insights to reach our motive.

B. Related Work

There are few authors who have worked into the hadoop projects and demonstrated as follows. A.Verma [1] models map and reduce phase individually through Hadoop MapReduce and spark. B.Aditya [2] gives his approaches to address the big data problems. W. wang et. al [4] was showcased hadoop with the balanced data but it fails to manage network traffic. S. Pandey [5] works on his paper to reduce the network blockage. They reduced the network traffic by introducing aggregation algorithm to send the different key, value pair to the reducer. C. Wang [6] explains how interrelation is important for two type of sales datasets and find the correlation between the brand sales and search volume. C. Pal [7] analysed the performance between the system and hadoop system with large datsets and also seen the behaviour of mapper and reducer tasks with amount of bytes written and read from the different sized files.

T.V. Kenekar [8] used MapReduce with FP-growth algorithm to improve the overall performance of FIM. J.Lin and C. Dyer [9] has introduced to join mapper to process multiple key value pair. Dhanya et.el. [10] has provided some improvements in the MapReduce model. The Hadoop MapReduce is the method which has master-slave attributes [11] as key/value k,v pairs.

C. Working on MapReduce model

MapReduce framework was firstly provided by google to deal with the huge datasets. Map Reduce on hadoop involves cluster and it makes parallel processing of data to shorten the elapsed time to run the algorithm. In this model, there are mainly classified into two functions. They are map method() and another one is reduce method(). With that MapReduce framework, users can stimulate their own code to get their desired output.

1) Mapper: Mapper method takes the input from the argument and applies the mapper function to all its incoming data as a key, value pair. Mapper then sort the key, value pair in the ascending order. Next step of the mapper is to divide the intermediate key and assigned it as proper key , value pair to the reducer for processing next stages of MapReduce.
2) Reducer: From the spliced output from the mapper is turned as input to the reducer. Reducer than suffle and sort the data into the same time to get the output from the reduce task as a file system.

Mapper Phase In MapReduce: The mapper phase of the MapReduce framework has been described below.

- Input is given the model.
- Map Model is taken the input in the form of key and value pair.
- The output pair is stored as a buffer in the memory.
- when the buffer occupies 80 percent of the data then it splits into the other disks.
- All the spill files are combined into a single large file which is partitioned and stored based upon the reducer.
- Mapper Input: Map(Keyin, valuein)
- Mapper Output: list(keyOut, IntermediateValue)

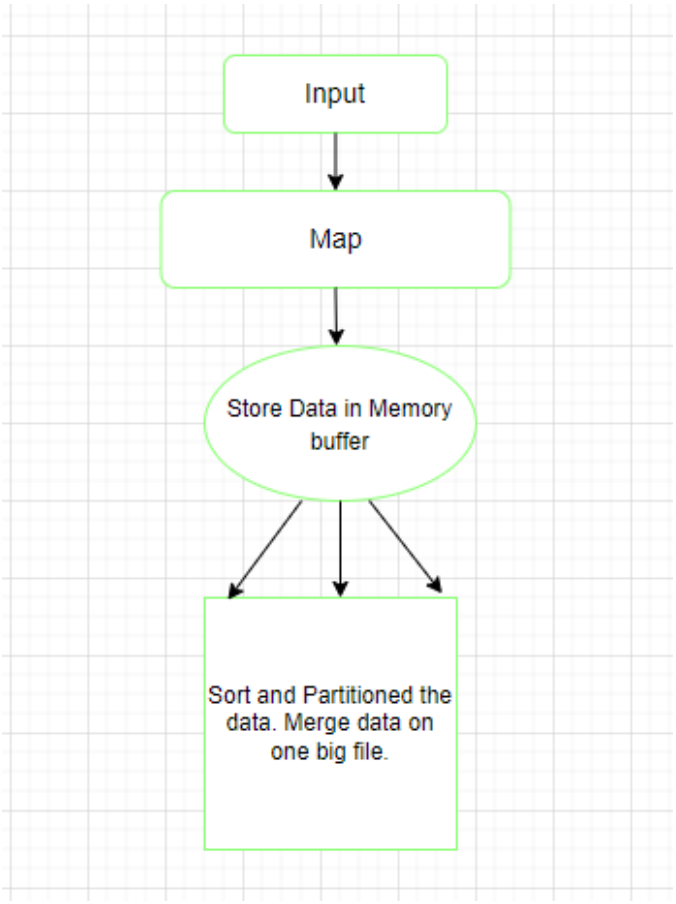


Fig. 3. Mapper Phase

Reduce Phase In MapReduce: The reducer phase of the MapReduce framework has been described below.

- The data from the map phase is given to the reducer and loaded into the memory.
- If the buffer attains 70 percent it is diverted into the different disk.
- Then the spilled data is then the merged and sort into the large files.
- Reducer method is initiated and reduce the file.
- Reducer Input: Reduce(KeyOut, list(intermediateValue))
- Reducer Output: list(OutValue)

Map Reduce components are given as below:

- Name Node - centrepeice of HDFS. It does not stores actual data.

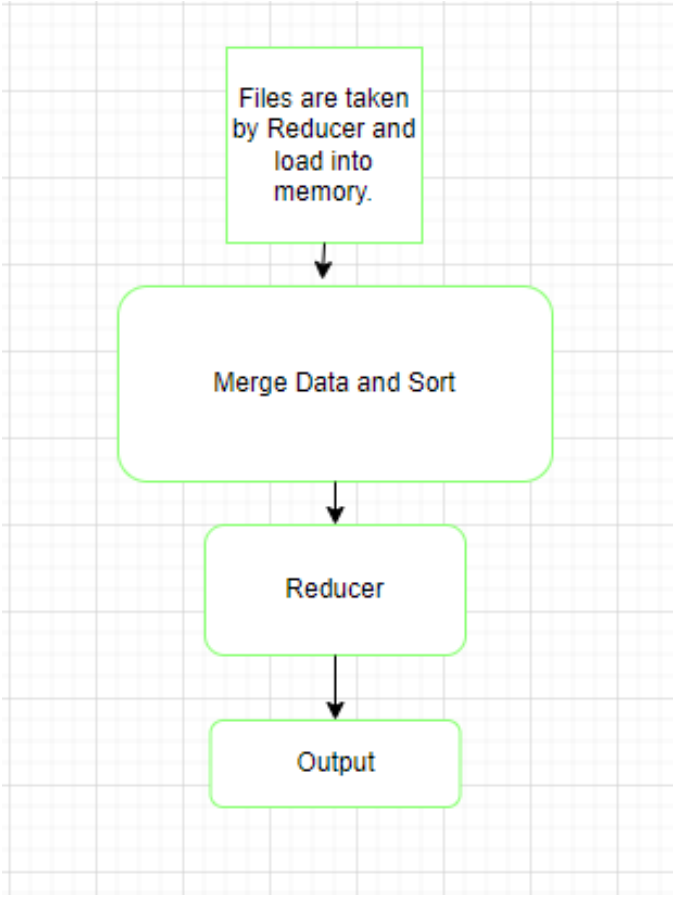


Fig. 4. Reducer Phase

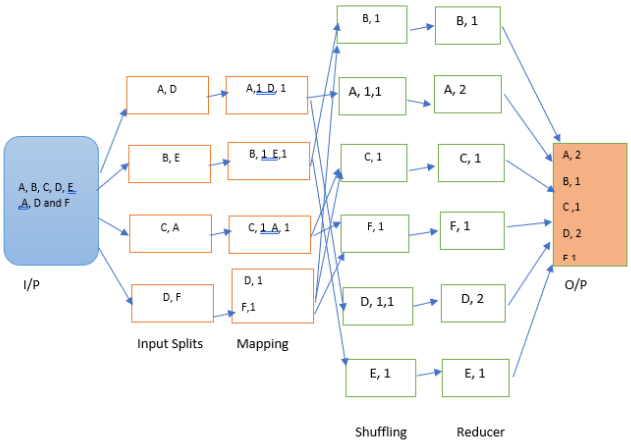


Fig. 5. Overall MapReduce Framework of BigData Hadoop

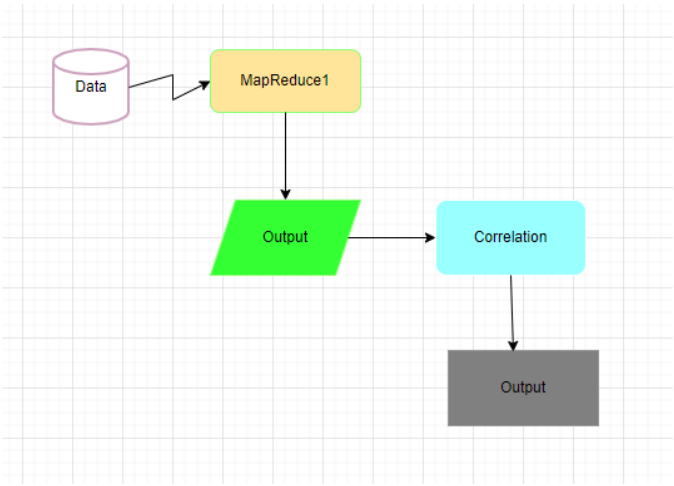


Fig. 6. Architecture Flowchart

- Data Node - keeps block level data of HDFS. The actual data is stored as slave node.
- Job Tracker - It is called as a master which creates and deploy the job.
- Task Tracker - It deploy the task and report to the Job Tracker. It is called as a slave.

The Overall working model of the MapReduce model is shown in the figure 5

IV. Implementation and Architecture

In this section, we will look into the application workflow, algorithm and then output is produced. With the retrieved output, we find the correlation matrix to find how the attributes are related each other. We will produce the several business insights to achieve our goal. Thus it can be analyzed through the visualization technology.

A. Architecture of the model

In this stage, the data will be inputted into the Map Reduce model1 to execute. Then the output generated by the first model will be given as input is transformed into the proper matrix to the correlation model. We can measure the interrelation between the attributes.

MapReduce algorithm has been implemented using java programming to make use of the map and reduce method to receive the desired output. MapReduce model has different inbuild framework to automate the implementation process for the huge data sets. With this package, the writing code would be easier. The below shown is the mapreduce algorithm of our model.

MAPREDUCE: Mapper1: Input- year, country, variable code, value

Mapper2: Input - year, direction, population

Output- Key - year value - (country, variable code, value, direction, population)

Reducer: In this section, value is retrieved as a string and make our logic to perform.

Output: Here we have retrieved two outputs to perform our analysis. value - (year, (country, value, direction, population)) value - (year, (variable code, value, direction, population))

Correlation: Overall ICT and population in all direction

value (value, population) Based on Employment and incoming people (value, population) Based on gross value added and outgoing people (value, population)

B. Visualization and insights

1) Predict how will be the European economic growth flows based on the people movement. We have taken two attributes from the output such as overall growth of the European countries and the population count of the Ireland. In this we can predict how will be the economic growth flow happens when the people movement from in/out to the countries. By the graph 7, we can say both the items are related to each other when population count goes up/down economy also acts upon the same.

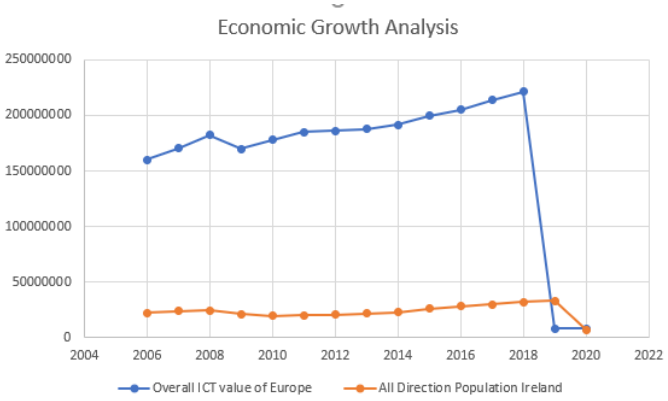


Fig. 7. Overall growth vs Population count

2) Based on gross value added of the sector, how much the country losses his own people. For this analysis, we have taken Gross value added of the UK and outward people who have moved out from the Ireland. This shows how both the variables are interact each other by the graph 8

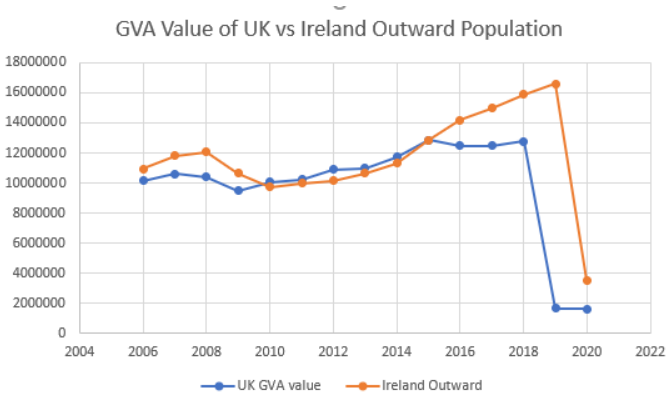


Fig. 8. UK Gva value vs Ireland Outward Population

3) Employee wages vs incoming people to the countries. We can predict how the incoming people is happening into the Ireland based on the employment value in Ireland. We have clubbed 3 group of year to analyse it in graph 9

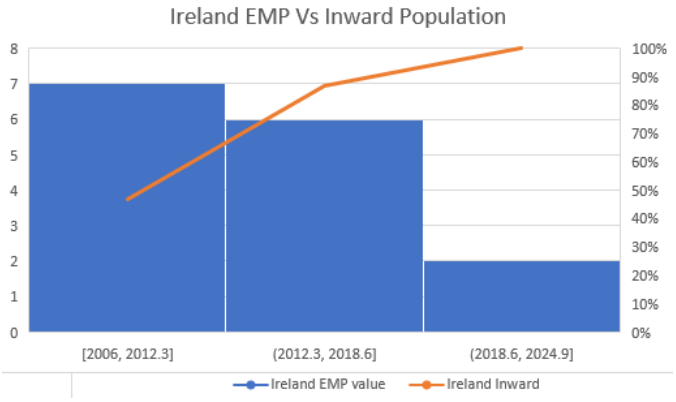


Fig. 9. Ireland Employment vs Inward population

4) In which the year, sector value(economic growth) and the population movement is more. By taking the interactive bar chart into account, we can see there is similar growth in both of the variables and there is highest point on the value of economic growth and population in the year 2018 and 2019 respectively.

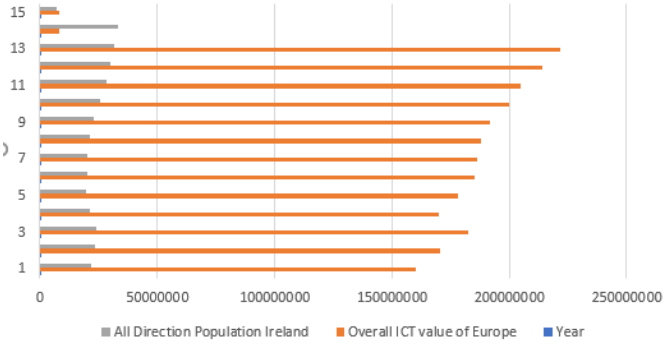
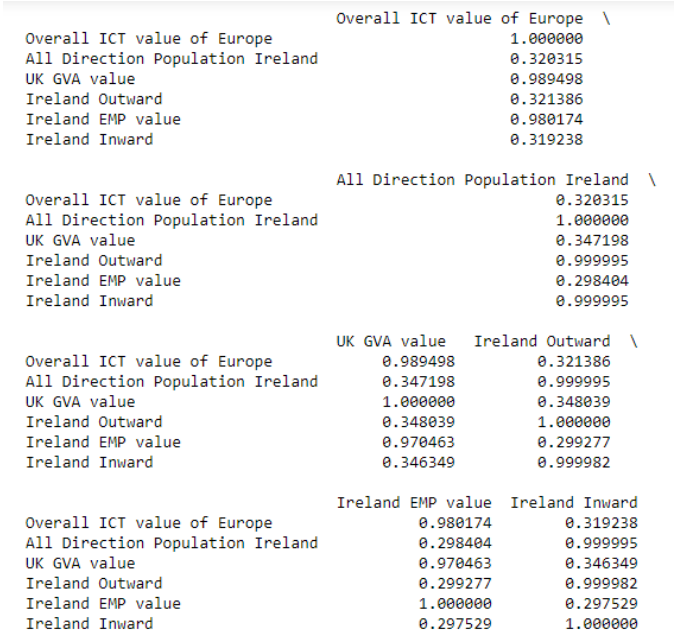


Fig. 10. Interactive Bar Chart

We can see there is connection between each attributes in all the 4 insights shown above. Thus the accuracy of the model is good by just seeing the analytical graph. We have analysed correlation by both Python and R. Below shown figure 11is from python output. The value close to the 1 are more related to each other and away from 1 are less related one.

V. RESULTS AND CONCLUSION

MapReduce Techniques were applied to the important dataset produced by government of Ireland and Europe. We were excited to use these Big datasets in this project to analyse economic growth and Travel Population using MapReduce concepts and visualize it to produce valuable insights. One of the surprising thing is to do data cleansing with java code and do analyse with mapreduce framework to produce the proper



	Overall ICT value of Europe
Overall ICT value of Europe	1.000000
All Direction Population Ireland	0.320315
UK GVA value	0.989498
Ireland Outward	0.321386
Ireland EMP value	0.980174
Ireland Inward	0.319238

	All Direction Population Ireland
Overall ICT value of Europe	0.320315
All Direction Population Ireland	1.000000
UK GVA value	0.347198
Ireland Outward	0.999995
Ireland EMP value	0.298404
Ireland Inward	0.999995

	UK GVA value	Ireland Outward
Overall ICT value of Europe	0.989498	0.321386
All Direction Population Ireland	0.347198	0.999995
UK GVA value	1.000000	0.348039
Ireland Outward	0.348039	1.000000
Ireland EMP value	0.970463	0.299277
Ireland Inward	0.346349	0.999982

	Ireland EMP value	Ireland Inward
Overall ICT value of Europe	0.980174	0.319238
All Direction Population Ireland	0.298404	0.999995
UK GVA value	0.970463	0.346349
Ireland Outward	0.299277	0.999982
Ireland EMP value	1.000000	0.297529
Ireland Inward	0.297529	1.000000

Fig. 11. Correlation Analysis Using Python

output. After correlation, we can see most of our insights are correlated to each other. So our aspect of thing is excellent to say as the best application. One of the key aspect of this result I found was the ICT value and Population keep on going in positive direction before 2019. After that due to covid-19 it drastically decreases. Thus the year 2018, ICT value is highest among all and in year 2019 travel in and out from the Ireland is increased. I assumed this might be people travelled to there hometown to stay with their families. This was the interesting insights. We can conclude by saying that nature and environment makes efficient changes to the economic growth and travel.

VI. FUTURE WORK

Future aim is to work on different hadoop concept and do more visualization to make an application which would be useful for people to check whether which country is best suited for employment, savings, education etc. Due to time constrains, this part is not covered in this project.

REFERENCES

[1] A. Varma, A. Hussein, and N. jain, "Big data management processing in hadoop map reduce and spark technology: A comparison," in *symposium on Colossal Data Analysis and Networking (CDAN)*, 2016.

[2] B. Aditya, M. birla, and U. Neir, "Addressing big data problem using hadoop and map reduce," in *Nirmala University International Conference on Engineering*, 2012.

[3] I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, "Big data quality framework: a holistic approach to continuous quality management," in *Journal of Big Data*, vol. 1, no. 115, 2021.

[4] W. Wang, K. Zhu, L. Ying, J. Tan, and L. Zhang, "Map task scheduling in mapreduce with data locality: Throughput and heavy-traffic optimality," in *INFOCOM, 2013 Proceedings IEEE, Turin*, vol. 1609-1617, 2013.

[5] S. Pandey, M. Supriya, and A. Shrivastava, "Aggregation algorithm to overcome data travel cost in mapreduce," in *International Conference on Computational Intelligence in Data Science*, 2017.

[6] C. Wang and C. Y. Hsu, "Rankings correlation study," in *IEEE International Conference on Big Data Analytics*, 2020.

[7] C. Paal, P. Agerwal, K. jain, and S. Agerwal, "A performance analysis of mapreduce task with large number of files dataset in big data using hadoop," in *Fourth International Conference on Communication System and Network Technologies*, 2014.

[8] T. V. Kenekar and A. R. Dani, "An efficient private fim on hadoop mapreduce," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 2016.

[9] J. Lin and C. Dyer, "Data-intensive text processing with mapreduce," in *Synthesis Lectures Human Language Technology*, vol. 3, no. 1, 2010, pp. 1–177.

[10] S. Dhaanya, M. Vysakan, and A. mahesh, "An enhancement of the mapreduce apriori algorithm using vertical data layout and set theory concept of intersection," in *Intelligent Systems Technologies and Applications*, vol. 385, 2016, pp. 225–233.

[11] B. Pratheeba. and M. Prathilothamai, "A study of distributed systems in realtime applications," in *International Journal of Control Theory and Applications*, vol. 9, no. 10, 2016, pp. 4233–4240.