**Dr B R Ambedkar National Institute of Technology, Jalandhar**



**INDUSTRIAL PRACTICAL TRAINING REPORT**

**Data Analyst Internship**

**at**

****

**For the award of the**

**Degree of Bachelor in Technology**

**in**

**Department of Computer Science & Engineering**

(Session: 2018-2022)

***Topic : Azure Data Platform for COVID Data Analysis for Health Provider Industry***

**Institute Supervisor: Submitted by:**

Dr. Urvashi Sanyamdeep Singh

(Assistant Professor) 18103083

Dept. of CSE B.Tech. CSE (Final year)

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Title** | **Page No.** |
| **1.** | **DECLARATION** | 3 |
| **2.** | **CERTIFICATE** | 4 |
| **3.** | **ACKNOWLEDGEMENT** | 5 |
| **4.** | **ABOUT THE COMPANY** | 6 |
| **5.** | **ABSTRACT** | 7 |
| **6.** | **WEEKLY OVERVIEW** | 8 |
| **7.** | **CHAPTER I - INTRODUCTION** | 9 |
| **8.** | **CHAPTER II - AZURE INTRODUCTION** | 11 |
| **9.** | **CHAPTER III - AZURE TECHSTACK** | 14 |
| **10.** | **CHAPTER IV - EXPLORATORY DATA ANALYSIS** | 18 |
| **11.** | **CHAPTER V - PRESENTATION TABLES AND VISUALIZATION** | 27 |
| **12.** | **CHAPTER VI – CONCLUSION & FUTURE SCOPE** | 30 |

**DECLARATION**

I hereby certify that the work which is being presented in this report is entitled, in fulfillment of the requirement for the award of Degree of **Bachelor of Technology** submitted in the department of **Computer Science and Engineering** of **Dr. BR Ambedkar National Institute of Technology, Jalandhar** is a record of my work carried out during Internship under the supervision of **Dr. Urvashi** (Assistant Professor CSE Dept., NIT Jalandhar) and under the mentorship of **Mr. Amit Kumar**. The matter presented in this report has not been submitted in part or full to any other University or Institute for the award of any degree.

No part of this report has been submitted elsewhere. All sources of knowledge used have been duly acknowledged.

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

**Signature of the supervisor**

**Dr. Urvashi**

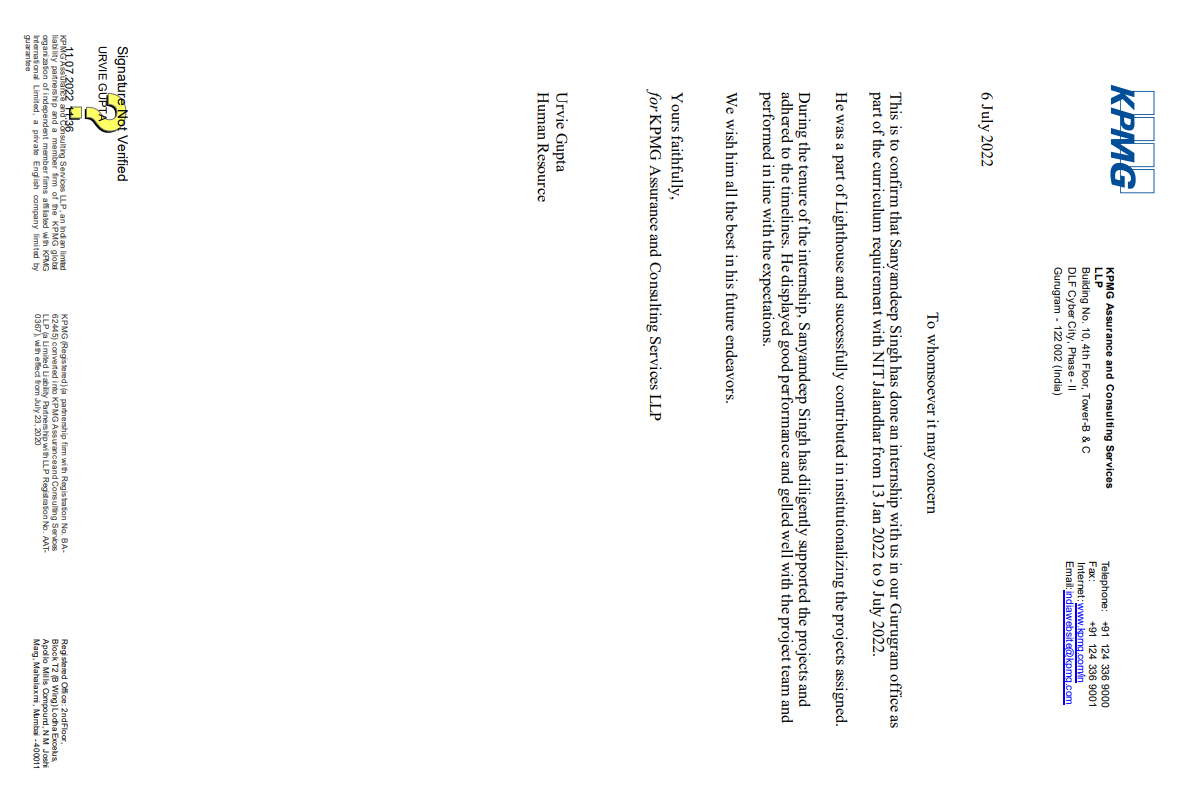
(Assistant Professor)

Dept. of CSE

Dr B R Ambedkar NIT Jalandhar

###### 

# CERTIFICATE



**ACKNOWLEDGEMENT**

This dissertation would not have been possible without the guidance and the help of several individuals who is one way or another contributed and extended their valuable assistance during the duration of the internship. I extend my supreme gratitude to the Training & Placement Cell and Computer Science and Engineering Department, NIT Jalandhar for organizing the whole internship program which has provided an opportunity to gain practical working experience.

I am extremely thankful to **Professor A. L. Sangal**, Head, Department of Computer Science & Engineering and my mentor Dr. Jagdeep Kaur, Assistant Professor who has provided all their direct and indirect support.

I would also like to express my special thanks to **Ms. Ridhima Handa**, HR, Head, of KPMG for giving me the opportunity to do internship within the organization.

I am thankful to **Mr. Amit Kumar, Partner**, KPMG and Shaista Khan, Assistant Manager, KPMG for devoting time and explaining how work is being done in the organization and providing required training.

I also like to express my gratitude to our families for their unwavering support and blessings throughout the process. Last but not least, I want to express our gratitude to all of our classmates for their unwavering support and fruitful talks over the duration of the project.

**ABOUT THE COMPANY**

KPMG provides risk, financial & business advisory, tax & regulatory services, internal audit, and corporate governance services and one of the Big Four accounting organizations. Through helping other organizations mitigate risks and grasp opportunities, KPMG can drive positive, sustainable change for clients, our people and society at large. KPMG entities have a domestic client base of over 2700 companies. It is a global approach to service delivery helps provide value-added services to clients. KPMG India has Global Centers of Excellence in the Telecom, Education, Energy, ERP, Sourcing and Business Excellence fields.

KPMG differentiation is derived from a rapid performance-based, industry-tailored and technology- enabled business advisory services delivered by some of the leading talented professionals in the country. KPMG professionals are grouped by industry focus and our clients are able to deal with industry professionals who speak their language. Our internal information technology and knowledge management system is responsible for the delivery of informed and timely business advice to clients.

My work was in the Lighthouse Department of the company.

The KPMG Lighthouse, Center of Excellence for Data, Analytics, and AI, combines the latest technologies and capabilities alongside our deep- rooted domain expertise to accelerate innovation and drive speed and relevance to our client’s businesses.

We leverage data, analytics, and artificial intelligence to accelerate digital transformation at a global scale.

**ABSTRACT**

This project presents the design of Azure Data Platform for ECPC COVID Data Analysis where we had to pull data from the ECDC website and create business insights for a Health Care Provider across European Countries. In this project, we used the technologies that we were taught throughout our training program.

Firstly, we identified all the azure services we would require to build the solution and designed the architecture, and provisioned them. Then, we created a pipeline to ingest the data in Azure Data Factory and stored it in the Data Lake. Next, the data was transformed and visualized using Azure Data Factory and Azure Data Bricks. At last, the created pipeline was triggered such that it would draw data at the end of the week automatically.

The visualizations generated showed the following results –

1. Smaller countries like Yemen had the highest Mortality Rate per case.
2. The no of cases per week was the most in mid-Sept and continually decreased thereafter.
3. France had the most no of hospital admissions during the pandemic.
4. Austria and Italy did most number of covid tests.
5. Countries like Lithuania, Netherlands, and Malta observed a higher positivity rate per test done.

**WEEKLY OVERVIEW**

|  |  |  |
| --- | --- | --- |
| Week | Time-Frame | Remarks |
| Week-1-  Week-2 | 13/Jan/2022– 15/Jan/2022 | * Introduction with Organization. * Laptop setup and softwares installation * Basic Courses on LinkedIn Learning. |
| Week-3-  Week-4 | 24/Jan/2022-  4/Feb/2022 | * Continue with LinkedIn Learning. |
| Week-5 | 7/Feb/2022-  11/Feb/2022 | * Training on Azure by CloudThat Team. * Introduction to Cloud Computing and Azure Fundamentals. |
| Week-6 | 14/Feb/2022-  18/Feb/2022 | * Introduction to Azure Data Fundamentals. * Basis of DBMS. |
| Week-7 | 21/Feb/2022-  25/Feb/2022 | * Introduction to Python Programming. |
| Week-8 | 28/Feb/2022-  04/Mar/2022 | * Introduction to Azure Data Factory and Other relevant topics. |
| Week-9 | 07/Mar/2022-  11/Mar/2022 | * Learn SQL Programming. * Introduction to T-SQL. |
| Week-10 | 14/Mar/2022-  18/Mar/2022 | * Introduction to Spark Programming. * Introduction to Hadoop – HDFS and Map Reduce. |
| Week-11-  Week-12 | 21/Mar/2022-  01/April/2022 | * Introduction to Azure Synapse Analytics. |
| Week-13 | 04/April/2022-  08/April/2022 | * Introduction to Azure Databricks and Cluster Creation. |
| Week-14 | 11/April/2022-  15/April/2022 | * Basic Introduction to Azure Log analytics |
| Week-15 | 18/April/2022-  22/April/2022 | * Basic Introduction to Azure DevOps * Introduce to Forage platform |
| Week-16 | 25/April/2022-  06/May/2022 | * Break and completion of first certification |
| Week-17–  Week-18 | 09/May/2022-  20/May/2022 | * 2-weeks break for college end semester exam and major project evaluation. |
| Week-19-  Week-20 | 23/May/2022-  31/May/2022 | * Consulting 101 Program. * Completion of Azure Certification. |
| Week-21-  Week-23 | 01/June/2022-  24/June/2022 | * Capstone Project |
| Week-24-  Week-25 | 27/June/2022-  08/July/2022 | * Forage platform tasks and 101 consultation program. |

**CHAPTER I** **- INTRODUCTION**

* 1. **INTRODUCTION TO THE PROJECT:**

As the increase in COVID cases across the world has set up a sense of distress among the global population, various countries have been largely affected by this economically. Countries have tried to make use of various response measures which have somewhat reduced the cases but still, the issue stands at large.

ECDC collaborates with national health protection agencies across Europe to enhance and develop pan-continental disease surveillance and early warning systems. ECDC develops authoritative scientific opinions regarding the risks posed by current and emerging infectious diseases by collaborating with health experts from across Europe.

Within the scope of its mandate, the Centre shall:

* + search for, gather, collect relevant, evaluate, and disseminate pertinent scientific and technical information;
  + provide scientific advice, and technical and scientific assistance, including training
  + provide adequate data to the Commission, the Member States, Community organizations and international organizations active in the public health field;
  + coordinate the European connectivity of bodies operating in the Centre's mission areas, including networks that emerge from public health departments supported by the Committee and operating he dedicated surveillance networks;
  + exchange information, expertise, and practice guidelines, and facilitate the creation and execution of joint actions.
  1. **OBJECTIVES OF THIS PROJECT:**

The main objectives of the presented work in this thesis are:

* + Covid data to be pulled, available at ECDC’s website, into Data Lake in its native format.
  + Analyze the columns and data type and transform accordingly across each of the datasets provided by the ECDC platform.
  + Bring out insights such as trends persisting to cases and deaths reported, weekly and daily hospital and ICU admissions/occupancy, etc.
  + The ingestion pipeline should be able to pull data automatically from ECDC’s website with a minimal number of linked services and datasets
  1. **CHALLENGES AND MOTIVATION:**

The main challenges and motivation of the work are:

* + To become aware of Microsoft Azure’s Services.
  + To implement the services to design a fully automated functioning pipeline.
  + To analyze the obtained results and draw insights through th

**CHAPTER II - AZURE INTRODUCTION**

Azure is Microsoft's cloud computing platform. It provides resources like computing, storage, network, big containers, databases, etc. over the internet. To set up our own data center, for example, we will need a significant amount of money, time, physical space, IT labor, etc. In such situations, Microsoft Azure becomes our savior. It makes our work easier by providing virtual machines, rapid data processing, analytical and monitoring tools, and so on. Azure's pricing is also transparent and economical. Pay-as-you-go, also known as "Pay only for what you use," means that you only pay for the services you use.

Azure will increase in size and quality in the coming days. Additional functions and tools are being added. Currently, there are two releases. Microsoft Azure v1 and Microsoft Azure v2 are well- known versions. Microsoft Azure v1 was JSON script-driven to a greater extent than its successor, Microsoft Azurev2, which features an interactive user interface for simplification and ease of use.

* **Capital**: Azure eliminates the high cost of hardware, so we do not need to be concerned about capital. You simply pay as you go and enjoy a cash-flow-friendly subscription-based model.
* Additionally, creating an Azure account is simple. Simply register on the Azure Portal and select the subscription you require to get started.
* **Less Operational Cost**: The operational costs for Azure are lesser since it works on its own servers whose sole purpose is to keep the cloud consistent and bug-free; it's also significantly more reliable than a server hosted on-premises.
* **Cost-Effective**: If we set up a server on our own, we must hire a tech support team to monitor it and ensure everything is running smoothly. Also, there may be instances where the technical support team is taking too long to resolve the server issue. Therefore, in this regard is far too inexpensive.
* **Easy Back Up and Recovery options**: Azure provides simple backup and recovery options for all your valuable data. In the event of a disaster, you can recover all of your data with a single click without affecting your business. Cloud-based backup and recovery solutions save time, avoid large up-front expenditures, and incorporate third-party expertise.
* **Easy to implement**: Implementing your business models in Azure is a breeze. With a few on-click operations, you are ready to go. There are even a number of tutorials to accelerate your learning and deployment.
* **Better Security**: Azure's safety is superior to that of local servers. Regard your sensitive data and business applications without concern. Since it is stored securely in the Azure Cloud. Even during natural disasters, when resources can be destroyed, Azure is a savior. The cloud is always dynamic.
* **Work from anywhere**: Azure gives you the flexibility to work from any location. Only a network connection and credentials are required. And because the majority of Azure cloud services offer mobile applications, you are not limited to the device you have available.
* **Increased collaboration**: With Azure, teams can access, edit, and share documents at any time and from any location. Together, they can achieve future objectives. Another advantage of Azure is that it maintains activity and data logs. Timestamps are one example of how Azure maintains records. Time stamps improve team collaboration by increasing accountability and establishing transparency.

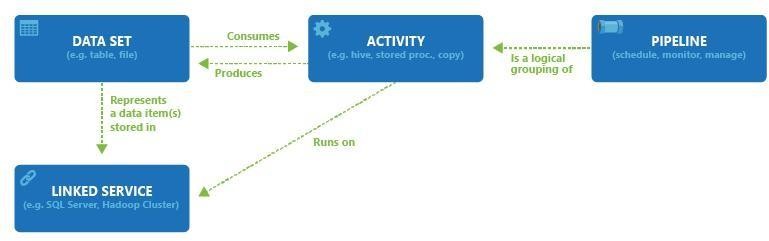
**Microsoft Azure Services:**

The following are the services Microsoft Azure offers-

* **Compute:** Contains Functions for serverless computing, Virtual Machines, Batch for containerized batch workloads, Virtual Machine Scale Set, Service Fabric for microservices and container orchestration, and Cloud Services for constructing cloud-based applications and APIs.
* **Networking:** Azure provides a number of networking tools, including the Virtual Network, which can connect to on-premise data centers; Load Balancer; Application Gateway; VPN Gateway; Azure DNS for domain hosting; Content Delivery Network; Traffic Manager; and Network Watcher for monitoring and diagnostics.
* **Storage**: It usually includes File, Queue, Disk Storage, and Blob, among others, as well as a Data Lake Store, Backup, and Site Recovery.
* **Web + Mobile**: Creating applications is a breeze due to the numerous application development and deployment services.
* **Containers**: Microsoft Azure also provides us a property that includes Container Service, which supports Kubernetes, Container Registry, and Docker Swarm, as well as tools for microservices.
* **Databases**: Azure also includes several SQL-based databases and related tools.
* **Data + Analytics**: Azure in addition to all has some major big data tools like HDInsight for Hadoop Spark and R Server.
* **AI + Cognitive Services**: Using Azure, developers can create applications with artificial intelligence features, such as the Computer Vision API, Bing Web Search, Face API, Video Indexer, and Language Understanding Intelligent.
* **Internet of Things**: It consists of IoT Edge and IoT Hub services that are combined with a variety of communications services and analytics.
* **Security + Identity**: Incorporates Security Center, Azure Active Directory, Key Vault, and Multi- Factor Authentication Services.
* **Developer Tools**: Includes c services such as Visual Studio Team Services, and Azure DevTest Labs to generate pipelines that help us

**CHAPTER III - AZURE TECH STACK**

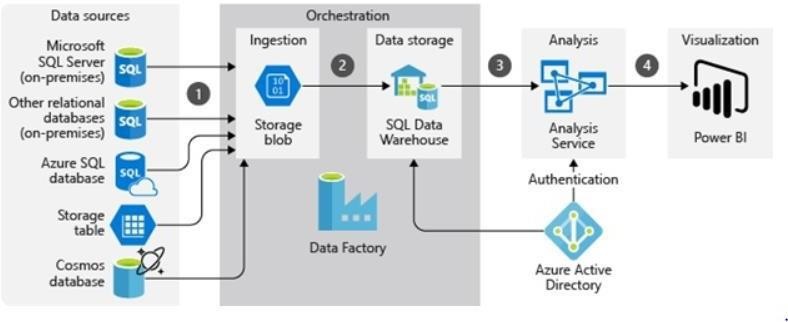
##### **AZURE DATA FACTORY:**

Azure Data Factory (ADF) is a Microsoft tool that allows you to move data from one location to another and optionally modify it as well. Occasionally, this activity is referred to as data engineering, ETL (Extract Transform Load), or ELT. SQL Server Integration Services is an older Microsoft product that also performs ETL operations (SSIS). They are distinct types of things with similar but not identical characteristics. SSIS is edited using on-premises tools like Visual Studio, whereas ADF is edited via a page in the Azure web portal.

***Figure 3.1.1 Azure Data Factory***

* 1. **AZURE SYNAPSE ANALYTICS**

Azure Synapse is an enterprise analytics service that expedites the time to insight across data warehouses and big data systems. Azure Synapse combines the best of SQL technologies used for enterprise data warehousing, Spark technologies used for big data, Data Explorer for log and time series analytics, Pipelines for data integration and ETL/ELT, and deep integration with other Azure services including Power BI, CosmosDB, and AzureML.



***Figure 3.2.1 Azure Synapse Analytics***

##### **AZURE DATABRICKS:**

Azure Databricks is a data analytics service offered by the Azure Cloud service platform. Azure Databricks offers two environments for the development of data-intensive applications, namely Azure Databricks SQL analytics and Azure Databricks. It is useful for those who wish to execute SQL commands on a data lake, generate multiple data visualizations in reports, as well as create and share dashboards.

It's used to facilitate coordination between Azure data engineers, developers, scientists, and machine learning engineers. The raw data is ingested into Azure via Azure Data Factory, Apache Kafka, Event Hub, or IoT Hub to create a big data pipeline. Azure data brick is utilized to retrieve data from numerous sources and transform it into groundbreaking insights using Spark.

Apache Spark is a framework for rapid cluster processing that is open-source and a well-known system for analyzing large amounts of data. This system measures data in a manner that contributes to performance enhancement. It is written in Scala, a high-level programming language, and supports APIs for Python, SQL, Java, and R.

#### **AZURE DATA LAKE GEN2**

Azure Storage becomes the foundation for building enterprise data lakes on Azure with Data Lake Storage Gen2. Data Lake Storage Gen2 was designed from the ground up to support multiple petabytes of data while sustaining hundreds of gigabits of throughput, enabling you to manage vast amounts of data with ease.

Data Lake Storage Gen2 now includes hierarchical namespaces as a fundamental its Blob storage. The hierarchical namespace organizes objects/files into a hierarchy of directories for efficient data access. A common object store naming convention mimics a directory structure with forward slashes. Data Lake Storage Gen2 facilitates this construction. Operations like renaming or deleting a directory are converted into atomic metadata operations. It is unnecessary to enumerate and process all objects that share the same name prefix as the directory.

Data Lake Storage Gen2 builds upon Blob storage and improves management, performance, and security as outlined below:

● It is not necessary to copy or transform data prior to analysis, thereby optimizing performance. The hierarchical namespace vastly improves the performance of directory management operations compared to the flat namespace on Blob storage, thereby enhancing the overall job performance.

● The ability to organize and manipulate files through directories and subdirectories simplifies management.

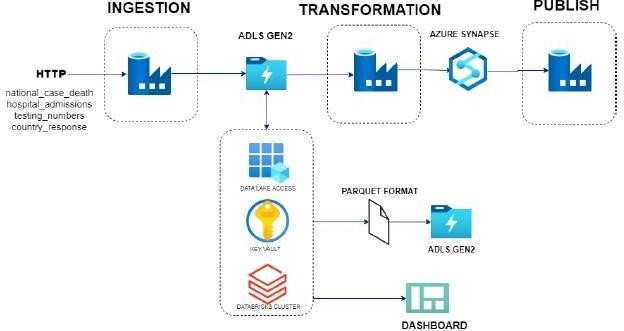
You can define POSIX permissions on directories or individual files, so security is enforceable.

###### AZURE KEY VAULT

Azure Key Vault is a cloud service for storing and accessing secrets in a secure manner. API keys, passwords, certificates, and cryptographic keys are examples of secrets that require stringent access control. Two types of containers are supported by the Key Vault service: vaults and managed hardware security module (HSM) pools. Software and HSM-backed keys, secrets, and certificates can be stored in vaults. Managed HSM pools only support keys that are HSM- backed. See the overview of the Azure Key Vault REST API for complete details. With the help of Azure Key Vault, we can securely store secrets such as passwords, API keys, and connection strings. Secrets usually are octet sequences that cannot exceed 25 KB in size.

Azure Key Vault is capable of storing both Cryptographic Keys (which are often used for encryption) and Azure Storage Account Keys. When Azure Storage Account keys are stored in Key Vault, the Key Vault will synchronize internally with the Storage Account and rotate the keys automatically.

* 1. **DATA ARCHITECTURE**



***Figure 3.6.1 Flow of Project***

The above figure shows the architectural flow of the project. Initially data from the ECDC website, majorly “national case death”, “hospital admissions”, “testing numbers” and “country responses” were pulled through the HTTP connector into Azure Data Lake Gen-2 account into a **“Raw”** data folder.

**CHAPTER IV - EXPLORATORY DATA ANALYSIS**

# DESCRIPTION:

* + 1. **SETTING THE CONTEXT:**

Before beginning a project, we must ensure that the data is suitable for modeling purposes. Exploratory Data Analysis (EDA) guarantees that Machine Learning data are prepared. In fact, the EDA makes surethe widespread use of data.

* + 1. **WHAT IS EXPLORATORY DATA ANALYSIS:**

Data scientists use Exploratory Data Analysis (EDA) to evaluate and investigate data sets and conclude their key characteristics, frequently employing data display methods. It aids in determining the optimal way to manipulate data sources to obtain the answers we require, making it easier for data scientists and us to identify patterns, ambiguous visions, theoretical or experimental tests, and so on.

EDA is primarily used to determine which data can be presented beyond the formal modeling or hypothesis testing and provides a better understanding of the data set dynamics and relationships. Additionally, it can be useful to determine if the mathematical methods being considered for data analysis are suitable. EDA strategies, which were created by the American mathematician John Tukey in the1970s, continue to be the most popular method of data acquisition today.

* + 1. **WHY EXPLORATORY DATA ANALYSIS?**

Prior to forming hypotheses, the EDA's primary function is to assist in analyzing that can aid in the identification of obvious errors, as well as the comprehension of data patterns, the identification of external factors or confusing events, and the discovery of interesting relationships between differences.

Data Scientists / Data Engineers / Data Analysts can use test analysis to ensure that the results they generate are valid and applicable to all desired business outcomes. Additionally, EDA assists participants by ensuring that they ask the appropriate questions. EDA can assist in answering questions concerning standard deviations, class variables, and confidence intervals. After EDA has been completed and the details have been drafted, its features can be utilized for more complex data analysis.

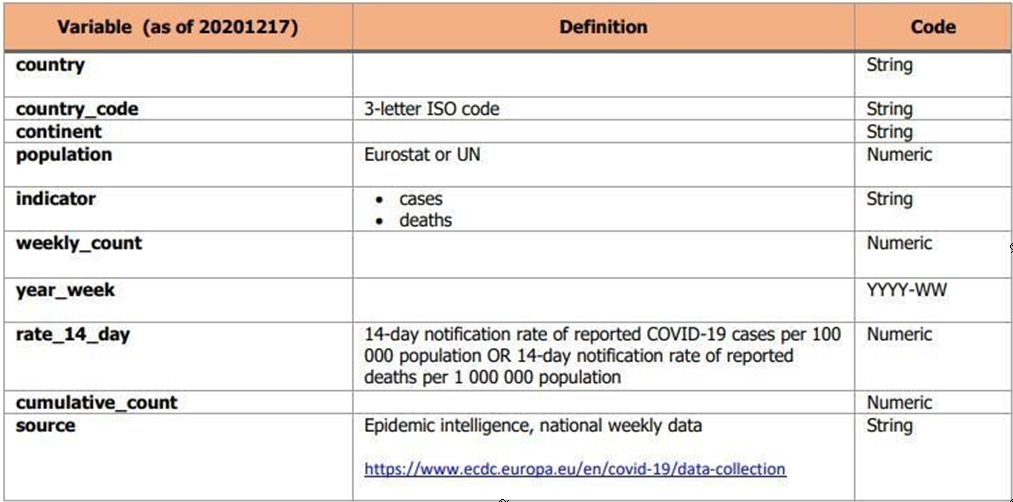
##### **PROBLEM STATEMENT:**

ABC Health is one of the largest hospital systems, with a chain of eleven hospitals, equipped with advanced biomedical instrumentation and a dedicated staff of 5000+ employees.

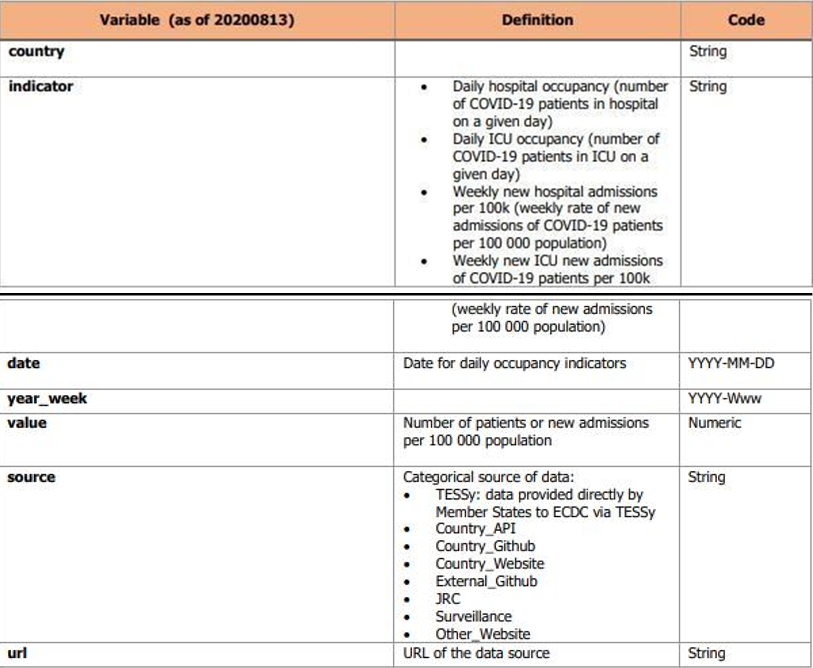
The company wants to have a dedicated team that can analyze data collected by ECDC Epidemic Intelligence to keep an eye on worldwide COVID trends such as reported cases, mortality rate, hospitalization & ICU admission rates & current occupancy, testing volume and country response which can act as feedback to improve their existing healthcare processes and patient monitoring system to respond optimally towards COVID.

They came to know about the seamless capabilities of Modern Data Platform Technologies that existwithin the ecosystem of Microsoft Azure and want to build an Azure Data Platform to bring insights out of the source mentioned to accelerate the response and services to save lives.

* 1. **UNDERSTANDING THE DATASETS:**
     1. **cases\_death.csv:** This file has data on the number of weekly cases and deaths of patients of various countries.

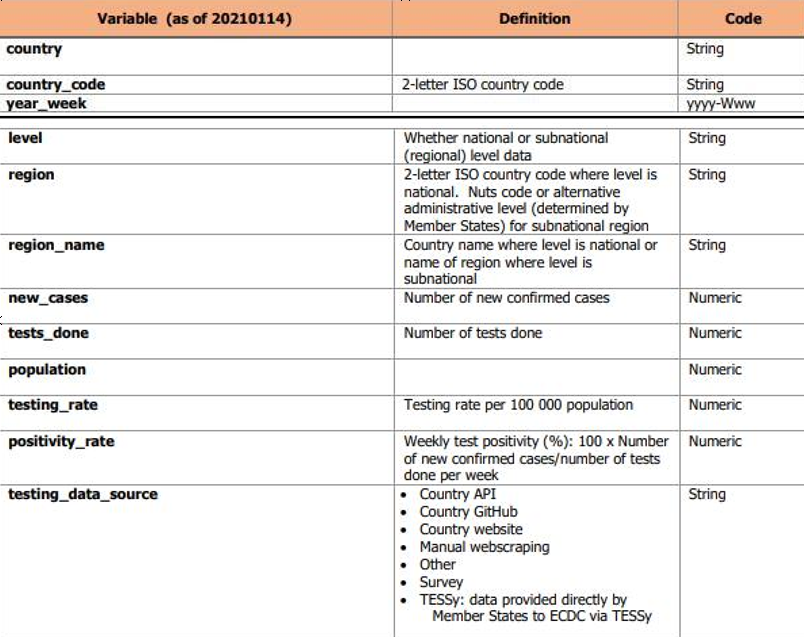
***Table 4.3.1 Deaths and Cases Dataset***

* + 1. **Hospital\_admissions.csv:** This file has daily data on the number of hospital & ICU admissions and weekly data on the number of new ICU & hospital admissions per 100000 population for various countries.



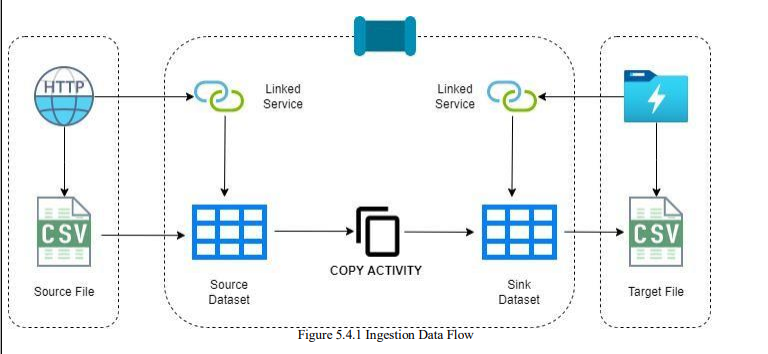
***Table 4.3.2 Hospital & ICU Admissions Dataset***

* + 1. **testing.csv:** This file has weekly data of new cases, no. of tests done, testing rates, and positivity rate of patients for various countries.



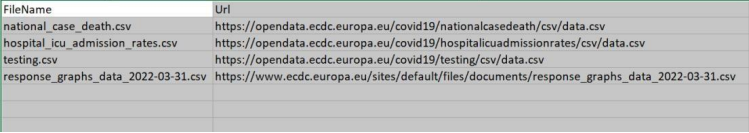
***Table 4.3.3 Testing Data Set***

* + 1. **country\_respones.csv:** This file has the data of response measures taken by each country to mitigate the spread of covid with the start date and end date.
  1. **INGESTION OF DATA:**



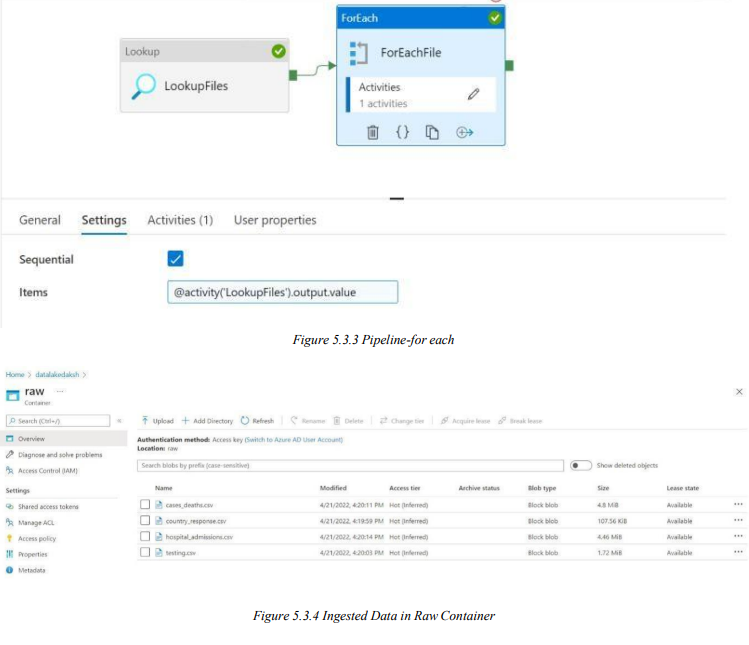
***Figure 4.4.1 Ingestion Data Flow***

A file containing URLs to the datasets is made available. The filenames are also made available as a different column, which makes it easier to name the files when copied into the raw container.

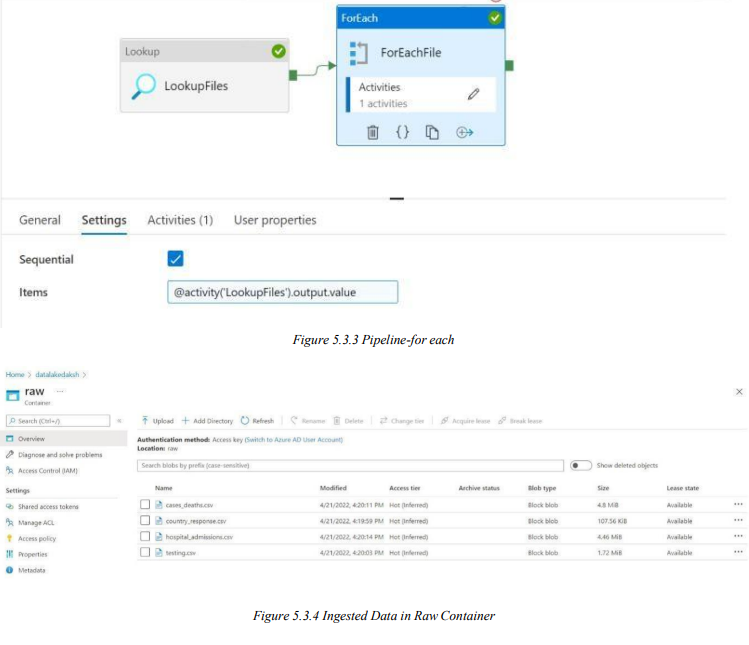


***Figure 4.4.2 File containing URLs***

A pipeline is created to copy the data from the ECDC website into Data Lake. A lookup activity is done on the file containing the URLs. A for Each Activity will look up the filename and URL and loop through each one of them to provide the necessary information to copy the activity. The copy data activity uses an HTTP connector to connect to the website to retrieve the data and sinks it to the raw layer in DataLake using the filename and URL as parameters. This reduces the number of datasets and linked services needed.



***Figure 4.4.3 Pipeline for each***

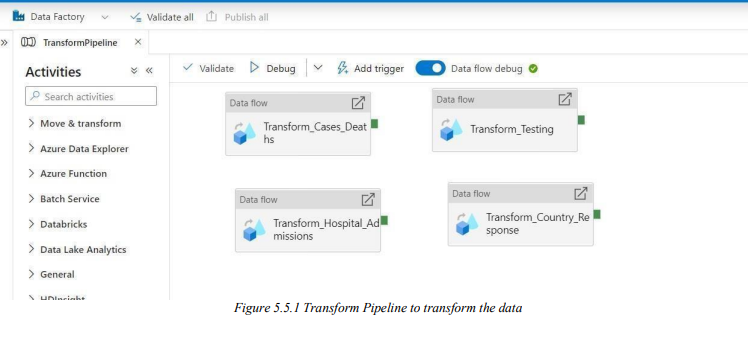


***Figure 4.4.4 Ingested data in each container***

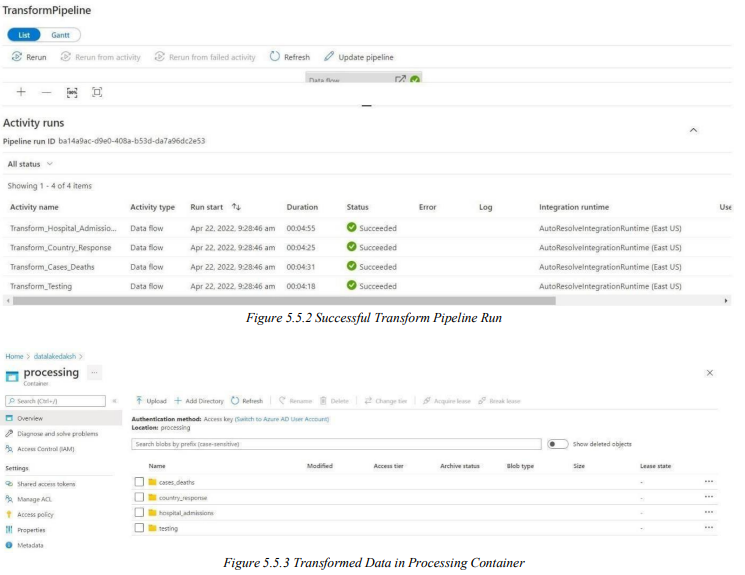
* 1. **TRANSFORMATION OF DATA:**

Azure Data Factory was used for Transformations in all the 4 CSV files. I have used 4 Mapping Data Flow activities and 4 different Data Flows parallelly in a single Transformation Pipeline. To start with the Source, have used the previously defined Dataset which was Sink in the Copy Activity as Source here for respective CSV files, and for Sink have created of 4 new datasets which will write the transformed data into the Processing Container in the Parquet format, since we required to create Managed tables using Databricks was done, that’s why we are using Processing container here instead of Processed container, which we will use to write the parquet files with saving them as table simultaneously. Because ADF doesn’t provide any activity to create Managed Tables separately.

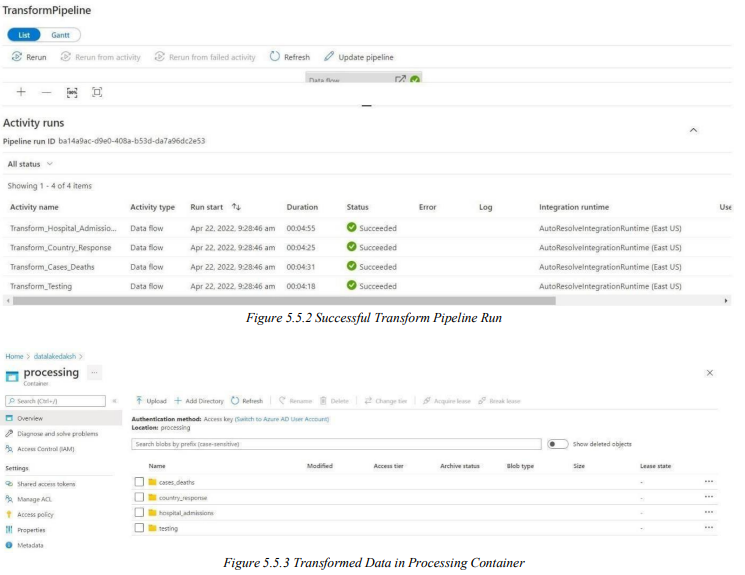
For each CSV file, the transformation Data flow is very much different. I have tried to set the data types of columns properly and also renamed some columns with the proper names.



***Figure 4.5.1 Transform pipeline to transform the data***



***Figure 4.5.2 Successful pipeline transform***



***Figure 4.5.3 Transformed data in processing container***

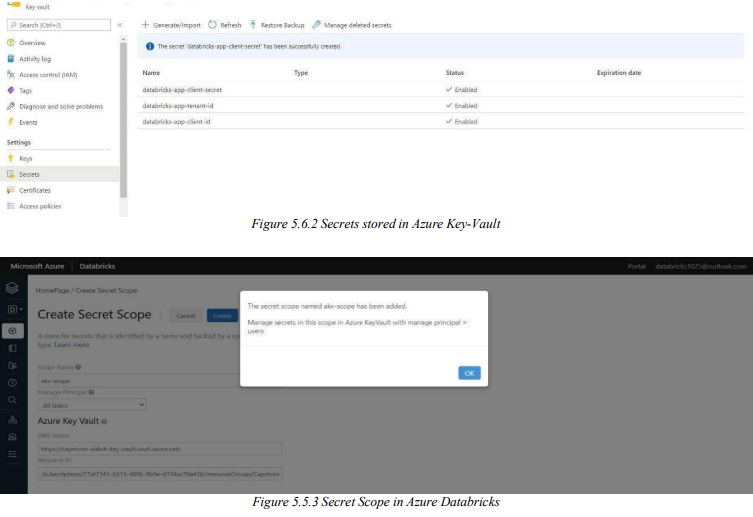
* 1. **CREATING EXTERNAL AND MANAGED TABLES:**



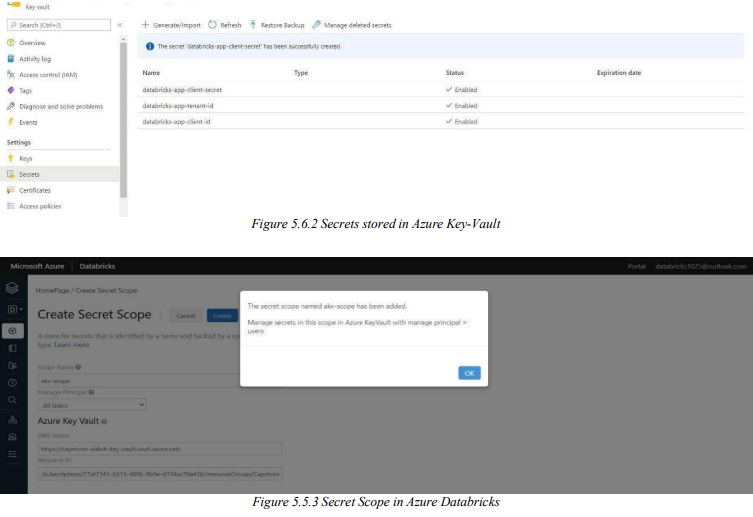
***Figure 4.6.1 Successful mounting and table creation pipeline run***

A single pipeline from Azure Data Factory namely, MountingTableCreationPipeline which contains 3 different notebook activity was used, which were linked using the Access token from the data bricks and using the linked service of Databricks from ADF. In this three-notebook activity, I have attached three different notebooks, one for mounting the ADLS container to Databricks, one for creation of Raw External Tables, and one for Processed Managed Tables.

For Mounting, the creation of an App in Azure Active Directory was done and registered and used Azure Key-Vault to store all the app secrets and used the secret scope when the values were required by the mounting code, i.e., I have used Key-Vault-backed secrets.



***Figure 4.6.2 Secrets stored in Azure key vault***

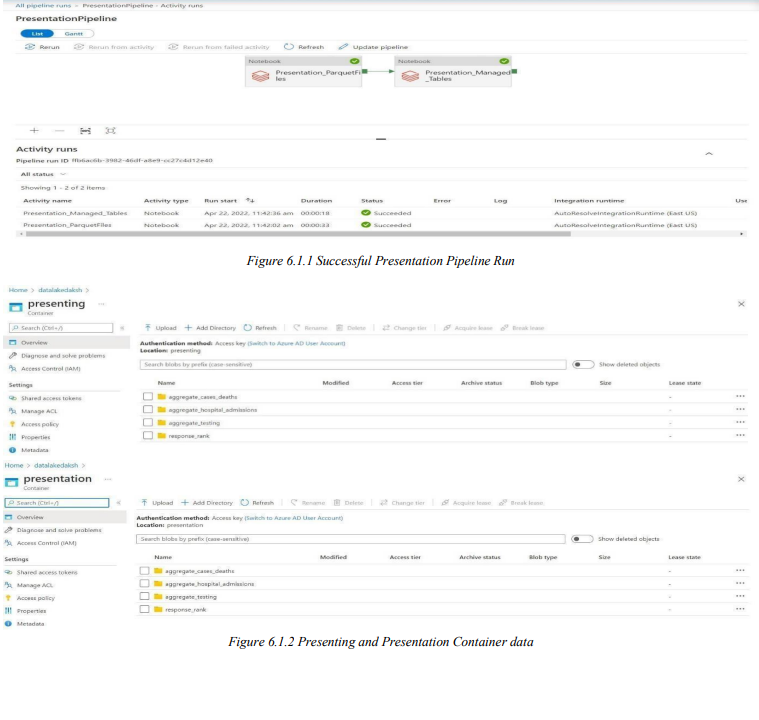


***Figure 4.6.3 Secret scope in Azure Databricks***

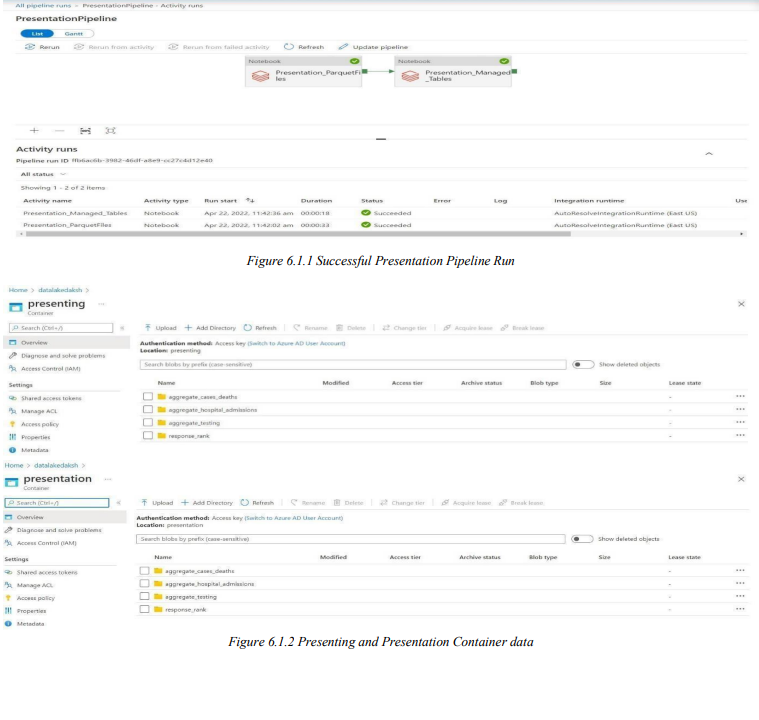
**CHAPTER V - PRESENTATION TABLES AND**

**VISUALIZATION**

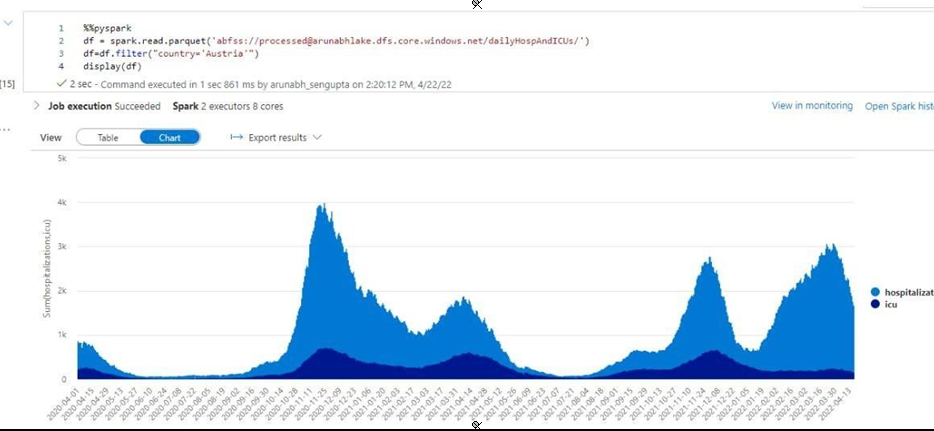
Here we have again used the Notebook activity to prepare the Presentation data and the Presentation Managed Tables to ease the task of the analysis further. Have used two notebooks. One for the creation of Presentation data and another one to create the managed tables on top of those presentation data.



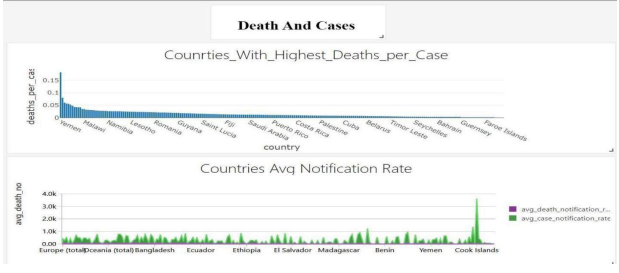
***Figure 5.1 Successful presentation of pipeline run***



***Figure 5.2 Presenting and Presentation Container data***

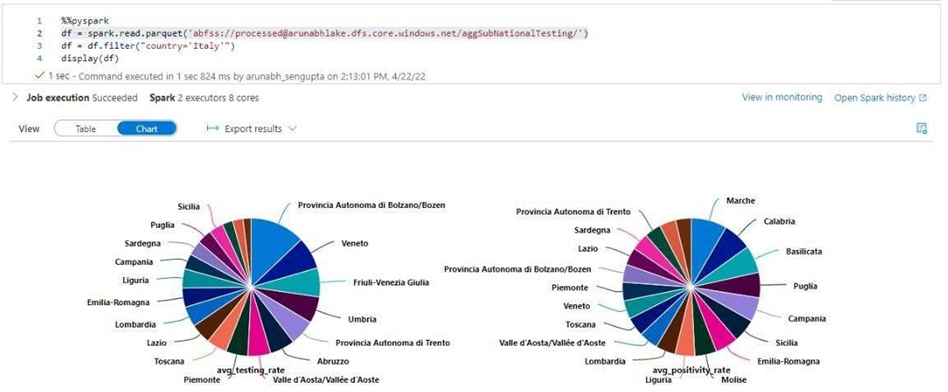


***Figure 5.3 Visualizing daily hospitalizations and ICUs for Austria***



***Figure 5.4 Visualizing countries with highest deaths per case notification***

.

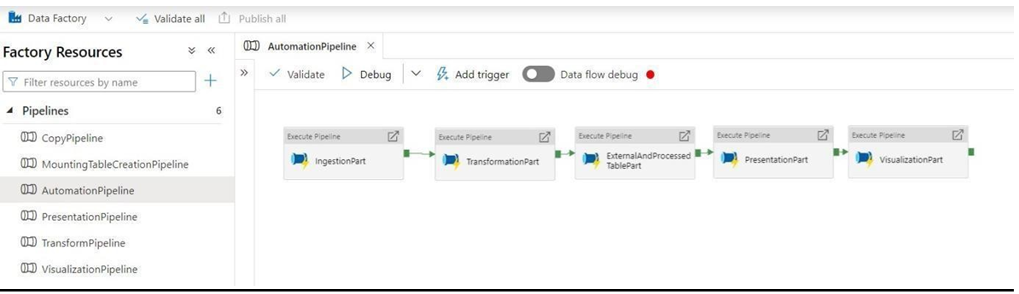


***Figure 5.5 Visualizing testing data for Italy***

**Automating Pipeline:**

The Complete process is Automated also using the Execution Activity and running the above 5 defined pipelines serially, but in this way, we can just set the schedule for weekly triggers only.

This can also be termed the master pipeline where the scheduled trigger is applied.



***Figure 5.6 Automation Pipeline***

**CHAPTER VI – CONCLUSION &**

**FUTURE SCOPE**

**6.1)** **CONCLUSION:**

* The architecture leading to the project, the resources used in the architecture are given and their description along with their use are explained.
* A complete procedure about how the data set was initially presented and how the data is ingested in the Data Lake has been discussed.
* Data transformation was implemented through Mapping Data Flow Activity as well as Azure Databricks. The resultant dataset has been shown.
* Dashboards and Visualization Results have been shown. Along with that, insights have been presented
* The resultant pipeline is scheduled and automated

**6.2) FUTURE SCOPE & APPLICATION:**

One very important future application is that the data we are provided with is a batch type pre sampled data which gets updated weekly, in order to make the process more efficient we could make use of data which gets updated daily or hourly hence help to make the decisions based on the data more quickly.

Another major applicable future scope could be the appliance of machine learning algorithms on the processed data set, this could help us define relationships much better as well as help us to predict future outcomes to some stage and hence be able to take necessary steps much before.

For the application part, we could make use of the same technology for various other systems as well, systems such as Movie Recommendation System or Optical Character Recognizer could be designed much more easily with the help of this pipeline structure.