IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An End To End Solution For Building A Data Platform For The Prediction And Reporting Of SARS-Covid19 Outbreak Using Azure Data Factory (ADF)

[1]K. Balakrishna Maruthiram, [2]Narayana Vijay Kumar,

[1]Assistant Professor of CSE, JNTU Hyderabad, [2]Master Student of JNTU Hyderabad

[1]Department of IT, JNTU Hyderabad, [2]Department of IT, JNTU Hyderabad

[1]JNTU Hyderabad, India, [2]JNTU Hyderabad, India

Abstract: In order to store and manage data from multiple sources, such as transactional systems, operational databases, and external sources, a centralized repository known as a data warehouse. It is an invaluable tool for organizations that need to make well-informed decisions based on large amounts of data. The structured and organized data stored in a data warehouse can store multidimensional information and facilitate easy retrieval and analysis, giving organizations valuable insights into their performance and operations. This makes strategic decsion making. The capacity to access both historical and present data, which enables companies to analyze trends and performance across time, is one of the main advantages of data warehousing. Moreover, changes in a data warehouse are usually planned for predetermined times, guaranteeing a constant and trustworthy supply of data for analysis. In business intelligence, data warehousing is essential because it enables firms to take calculated actions that lead to goal achievement.

1.INTRODUCTION

Azure Data Factory is Azure's cloud ETL service for scale-out serverless data integration and data transformation. It offers a code-free UI for intuitive authoring and single-pane-of-glass monitoring and management. You can also lift and shift existing SSIS packages to Azure and run them with full compatibility in ADF. SSIS Integration Runtime offers a fully managed service, so you don't have to worry about infrastructure management. This makes ADF an ideal solution for organizations looking to scale their data processing and transformation operations in the cloud.

Massive volumes of data may be processed fast and simply using ADF's server-less design, which eliminates the need for complicated setup or administration. ADF is a robust and adaptable solution for data integration and transformation. It also provides a wide range of data sources and formats with support, as well as a number of security and compliance capabilities to assist enterprises fulfill their data protection needs. Azure Data Factory (ADF) is a state-of-the-art solution for data integration and transformation that gives businesses the capacity to efficiently handle and examine vast volumes of data from multiple sources. Specifically engineered for cloud-based operations, ADF offers the scalability and flexibility required by enterprises to remain on the cutting edge of data processing and transformation. Organizations can handle large volumes of data fast and simply with its server-less design, which eliminates the need for complicated setup or configuration. One of the main advantages of utilizing ADF is that users can easily build and maintain data pipelines thanks to its user-friendly, code-free user interface. This makes it ideal for organizations looking

to streamline their data processing and transformation operations, as it eliminates the need for complex coding and manual configuration.

Furthermore, ADF provides single-pane-of-glass monitoring and management features that make it simple for businesses to keep tabs on the effectiveness and health of their data pipelines in real time. Compatibility with current SSIS packages, which can be simply lifted and moved to Azure for complete compatibility, is another advantage of ADF. This enables businesses to use their current resources and expertise while leveraging their investments in data integration and transformation. Lastly, ADF offers a number of security and compliance capabilities, including support for compliance with several industry standards and laws, encryption both at rest and in transit, to assist enterprises in meeting their data protection obligations. ADF is a flexible data format that can handle a broad variety of data sources.

2.OBJECTIVE

The purpose of this project is to use Azure Data Factory (ADF) to build a data pipeline that can be applied in real-world settings. Several sources, including the World Health Organization (WHO) and the European Centre for Disease Prevention and Control (ECDC), will provide the data. After that, ADF will be used to feed the data into Azure Data Lake Gen2, a cloud-based data storage service. Several big data services will be employed to evaluate and transform the data after it has been deposited in the data lake. These services will aid in sorting through the copious volumes of data and gleaning insightful information that may guide choices and spur advancement.

The pipes built as part of this project will be constantly monitored to make sure everything is operating as it should. Azure Monitor, Log Analytics, and ADF will be used to do this monitoring. With the use of these technologies, any possible concerns or difficulties that can surface during pipeline operation can be quickly identified and resolved. In conclusion, the goal of this project is to show how to use Azure Data Factory (ADF) to construct a real-world data pipeline. The pipeline will collect data from various sources, store it in a data lake hosted in the cloud, use big data services to analyze and process the data, and keep an eye on the pipelines to make sure they are operating properly.

3.REQUIREMENTS

The requirements for the project to develop a data factory for COVID-19 analysis will vary depending on the specific goals and objectives of the project. However, some common requirements that may need to be considered include:

The first and foremost we need to have a Azure subscription, either free trial or a paid version. All the resources of theproject are utilised from the portal.

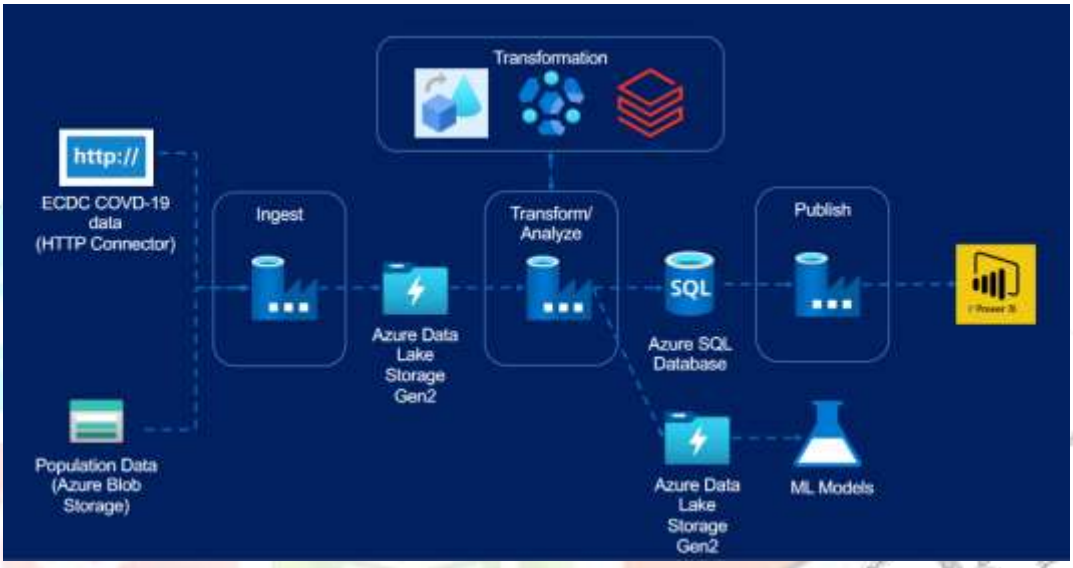
1.**Data Sources:** To build a comprehensive data pipeline, it is important to identify and gather data from relevant sources, such as ECDC and other health organizations, to analyze and make predictions about the spread of the virus.

2.**Data Storage**: The data collected from various sources will need to be stored in a centralized repository, such as Azure Data Lake Gen2, which provides the scalability and reliability required for large-scale data processing.

3.**Data Integration and Transformation**: To prepare the data for analysis, it may be necessary to perform various integration and transformation operations, such as data cleaning, aggregation, and normalization. ADF can be used to automate these processes, allowing for quick and easy data preparation.

4. **Monitoring and Management**: To ensure that the data pipeline is functioning properly, it will be important to implement monitoring and management tools, such as Azure Monitor and Log Analytics, to monitor the performance and health of the pipeline and ensure that data is being processed correctly.

4.PROPOSED SYSTEM:



A high-level system to allow COVID-19 response is suggested for the Azure Data Factory project:

1.Data Collection: Gathering information on COVID-19 is the initial stage, and it may be done from a variety of sources, including the World Health Organization (WHO), the ECDC, and other pertinent sites. This information may be in CSV, JSON, or other forms.

2. Data Ingestion: The data must be ingested into a data storage, such as Azure Blob Storage or Azure Data Lake Gen2, after it has been collected. This technique may be made scalable and automated with Azure Data Factory.

3. Data Transformation: Putting the data into a format that can be analyzed is the next stage. This might entail aggregating and summarizing the data in addition to cleaning and modifying it.This transformation may be carried out via Azure Data Factory with Azure HDInsight, Azure Databricks, or Azure Stream Analytics.

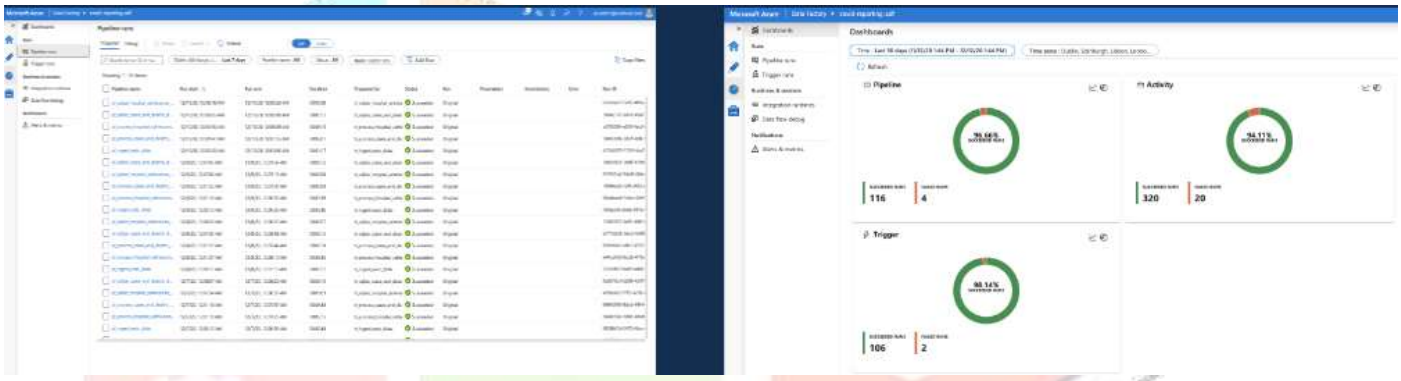
4. Data analysis: After the data has been converted, machine learning models may be used to analyze it further and draw conclusions and predictions. Azure Databricks and Azure Machine Learning may be used for this.

5. Data Visualization: Putting the analysis's findings and conclusions into visual form is the last stage. Power BI and other Azure-integrable visualization tools can be used for this.

6.Observation and Administration: A single pane of glass for pipeline administration and monitoring is offered by Azure Data Factory. The pipeline may be managed and kept an eye on to make sure it is operating smoothly and effectively using Azure Monitor and Log Analytics.

5. RESULTS :

COVID-19 Prediction and Reporting



6. APPLICATIONS:

1. **Fraud detection**: One crucial instrument for maintaining the integrity of financial transactions is machine learning. The model employs algorithms to find trends and abnormalities that might point to fraudulent behavior after being trained on a sizable quantity of data. By highlighting possibly fraudulent transactions for a human analyst to examine, this helps lower the possibility of fraud-related losses.

2. **Diagnostics in healthcare**: Machine learning has the ability to completely change how illnesses are identified and managed in the healthcare sector. The burden on healthcare systems is rising as a result of the explosion in global population and the growth in illnesses linked to a poor lifestyle. Large volumes of medical data may be analyzed using machine learning algorithms, which enable healthcare professionals to diagnose patients more quickly and accurately.

3. **content prediction.:** Machine learning models may anticipate a variety of future events by evaluating vast quantities of data. These predictions include which age groups may be afflicted by an illness, which symptoms may be common, mortality rates, hospitalization information, and much more. By guiding public health policy and resource allocation, these data can lessen the effect of illnesses and other hazards to public health.

4. **Equipment maintenance**: Machine learning has the ability to dramatically enhance how businesses manage their assets in the field of equipment maintenance. Machine learning algorithms can detect patterns and trends in historical maintenance activity data, which may be used to estimate the probability of equipment failure. By using this data to plan preventive maintenance, you can lower the risk of unplanned downtime and repair costs. Maintenance teams can save money by replacing or repairing

equipment before it breaks, for instance, if a machine learning model indicates that a certain piece of equipment is likely to fail soon. This helps to save unplanned downtime expenses.

7.CONCLUSION:

Azure Data Factory is a highly versatile and scalable cloud-based data integration service that enablesorganizations to seamlessly integrate and transform both cloud and on-premises data. Its user-friendly interface, combined with its ability to handle complex data transformations, makes it a valuable tool in any data platform and machine learning project. In the context of the COVID-19 pandemic, machine learning has been widely used to predict the spread of the virus and to forecast the cumulative number of cases. Azure Data Factory, with its low-cost and highly available data integration capabilities, plays a critical role in this effort. By combining the power of machine learning and Azure Data Factory, organizations are able to gather and analyse vast amounts of data to make more informed predictions and better respond to the challenges posed by the pandemic.

8.REFERENCES:

[1] R. Niehus, P. Martinez de Salazar Munoz, A. Taylor, M. Lipsitch, Quantifying bias of COVID-19prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers (2020).

[2] P.Pulla, Covid-19: India imposes lockdown for 21 days and cases rise, 2020.

[3] O. Analytica, Japan'S partial COVID-19 lockdown may be insufficient, Emerald Expert Briefings(oxanes).

[4] J. Thornton, Covid-19: A&e visits in england fall by 25% in week after lockdown, 2020.

[5] Y. Zhang, B. Jiang, J. Yuan, Y. Tao, The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: a data-driven SEIQR model study, medRxiv (2020).

[6] ECDC Website for Covid-19 Data —https://www.ecdc.europa.eu/en/covid-19/data

[7] Euro Stat Website for Population Data -https://ec.europa.eu/eurostat/estat-navtreeportlet-prod/BulkDownloadListing?file=data/tps00010.tsv.gz

[8] https://azure.microsoft.com/en-in/products/data-factory/