VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND INFORMATION SYSTEM

J-COMPONENT

TOPIC – TOKYO OLYMPICS DATA ANALYSISUSING AZURE SYNAPSE ANALYSIS

SWE2011- Big Data Analytics

SLOT – C2+TC2

Guided by-

SATHIYAMORTHY E.

By-

Kumar Shivam (20MIS0116)

Shashwat (20MIS0255)

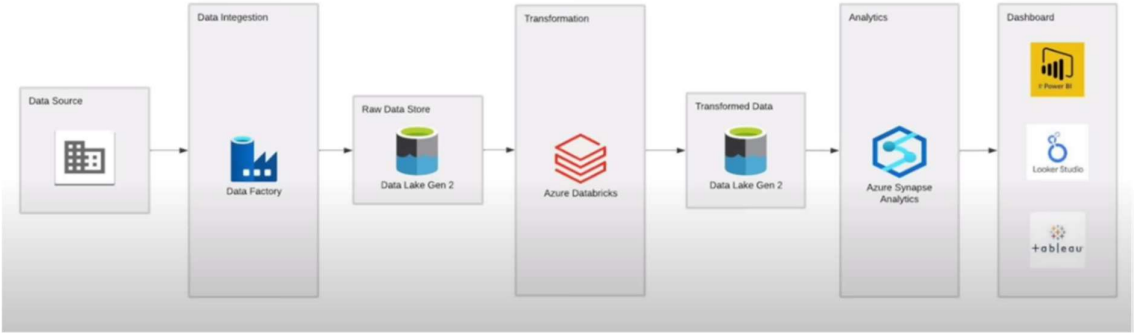
Logesh V. (21MIS0457)

ABSTRACT –

This project focuses on conducting comprehensive data analytics on Tokyo Olympics data through a systematic and efficient cloud-based approach utilizing Azure services. Beginning with data collection from diverse sources. Data integration will be seamlessly orchestrated using Azure Data Factory, ensuring a smooth flow of information. The raw data will then be stored in a Data Lake, leveraging AWS services.

For data transformation, Azure Databricks, a powerful Apache Spark-based platform, will be utilized to preprocess and clean the data before storing the transformed dataset back into the Data Lake. The subsequent stage involves leveraging Azure Synapse Analytics, formerly known as SQL Data Warehouse, for large-scale analytics and querying. The insights derived from this analytics process will be visually represented using Power BI, creating interactive and informative dashboards.

Throughout the project, a keen focus will be placed on security, compliance, and documentation. Security measures will be implemented to preserve data confidentiality, and compliance with data governance policies will be maintained. The project aims to provide a structured framework for end-to-end data analytics, from data collection to visualization, leveraging the strengths of Azure services for optimal results.



TERMINOLOGIES

Azure Data Factory (ADF): Azure Data Factory is a cloud-based data integration service provided by Microsoft Azure. It enables users to create, schedule, and manage data pipelines, facilitating the movement and transformation of data from various sources to destinations. ADF supports both structured and unstructured data and offers a visual interface for designing, monitoring, and managing data workflows. With connectors to various data sources such as Azure Blob Storage, Azure SQL Database, and on-premises sources, ADF allows organizations to streamline data movement and processing tasks, facilitating data-driven decision-making.

Azure Data Lake Gen2: Azure Data Lake Storage Gen2 is a scalable and secure data lake solution built on Azure Blob Storage. It provides capabilities for storing large volumes of structured and unstructured data, including files, images, videos, and logs. Data Lake Gen2 offers features such as hierarchical namespace, fine-grained access control, and high throughput for processing big data workloads. With integration with Azure services like Azure Databricks and Azure Synapse Analytics, Data Lake Gen2 enables organizations to store, manage, and analyze data at scale, empowering data-driven insights and decision-making.

Azure Databricks: Azure Databricks is a unified analytics platform provided by Microsoft Azure in collaboration with Databricks. It combines the power of Apache Spark with a collaborative workspace, enabling data engineers, data scientists, and analysts to work together on big data and machine learning projects. Azure Databricks provides a scalable and secure environment for processing and analyzing large datasets, with features such as interactive notebooks, job scheduling, and built-in machine learning libraries. By leveraging Spark's distributed computing capabilities, organizations can accelerate data processing and derive valuable insights from their data.

Azure Synapse Analytics: Formerly known as Azure SQL Data Warehouse, Azure Synapse Analytics is a cloud-based analytics service that brings together enterprise data warehousing and big data analytics capabilities. It allows organizations to query and analyze data from various sources, including data warehouses, data lakes, and streaming data, using familiar tools and languages like T-SQL and Apache Spark. Synapse Analytics offers features such as data integration, data warehousing, big data analytics, and machine learning, enabling organizations to derive insights from their data and drive business outcomes.

Power BI: Power BI is a business intelligence and analytics tool developed by Microsoft. It allows users to visualize and analyze data through interactive dashboards, reports, and data visualizations. Power BI connects to a wide range of data sources, including databases, cloud services, and on-premises sources, enabling users to create insightful visualizations and gain actionable insights from their data. With features such as natural language query, AI-powered insights, and collaboration capabilities, Power BI empowers organizations to make data-driven decisions and share insights across the organization

Github: GitHub is a web-based platform utilizing Git for version control. It hosts repositories for code management, enabling tracking of changes and collaboration. Git facilitates efficient codebase management through branching, merging, and conflict resolution. Collaboration tools like pull requests and issues aid in code review and task tracking. GitHub fosters community engagement, enabling contribution to open-source projects. It integrates with various tools and services for automation and extends functionality through a marketplace.

LITERATURE SURVEY

1.

TOPIC: Azure Synapse Analytics Use Cases and Reference Architecture

ABSTRACT: This final chapter of the book concludes our exploration of Azure Synapse Analytics. Throughout the preceding chapters, we have delved into various aspects of this powerful platform, providing you with a solid foundation for embarking on your journey with Azure Synapse Analytics. Given its complexity as an integration of multiple tools and technologies, understanding its architecture and core components can be challenging. To address this, we have dedicated separate chapters to each of these core components, including Synapse SQL, Synapse Spark, Synapse Pipeline, Synapse Link, Synapse Workspace, and Synapse Studio. Our aim has been to provide detailed insights into each component, facilitating a deeper understanding of Azure Synapse Analytics as a whole.

2.

TOPIC: Hyperspace: the indexing subsystem of azure synapse

ABSTRACT: This paper discusses Microsoft's recent introduction of Azure Synapse Analytics, which provides a seamless integration across data ingestion, storage, and querying utilizing Apache Spark and T-SQL for data stored in the lake, encompassing files and warehouse tables. The focus of this paper is on Hyperspace, the indexing subsystem underlying Synapse, detailing its design and implementation. Hyperspace allows users to create various secondary indexes on their data, manage them through a multi-user concurrency model, and automatically utilize them for query acceleration without necessitating any changes to their application code. The development of Hyperspace has been influenced by feedback from numerous enterprise customers. The paper covers Hyperspace's design principles, user-facing APIs, concurrency control mechanism for index access, index-aware query processing techniques, and maintenance procedures for handling index updates. Evaluations conducted using standard industry benchmarks and real customer workloads demonstrate that

Hyperspace can significantly accelerate query execution, by up to 10 times in certain scenarios, and even by orders of magnitude in specific real-world workloads.

3.

TOPIC: PolyBase in Azure Synapse Analytics

ABSTRACT: After exploring the array of new data sources facilitated by PolyBase V2, our focus shifts in this chapter to examine PolyBase's functionality for ETL tasks within Azure SQL Data Warehouse. We begin with an overview of PolyBase's capabilities and how Azure SQL Data Warehouse introduces complexities in functionality. Subsequently, we proceed to create a new database in Azure SQL Data Warehouse. Utilizing PolyBase, we demonstrate the process of reading data from Azure Blob Storage into Azure SQL Data Warehouse, along with providing insights on optimizing data loading performance. Additionally, we delve into the unique advantages offered by PolyBase when integrated with Azure SQL Data Warehouse.

4.

TOPIC: New query optimization techniques in the Spark engine of Azure synapse

ABSTRACT: The primary expenses incurred during the execution of big-data queries stem from stateful operators, such as sort and hash-aggregate, which often require intermediate data materialization in memory, as well as exchanges that entail data materialization to disk and data transfer over the network. This paper focuses on various query optimization techniques aimed at mitigating the costs associated with these operators. Initially, we present a novel exchange placement algorithm that surpasses existing methods by minimizing the amount of data exchanged while maximizing computation reuse through multi-consumer exchanges. Additionally, we introduce three partial push-down optimizations designed to push partial computation derived from existing operators (e.g., group-bys, intersections, and joins) below stateful operators. While these optimizations are broadly applicable, we observe that two of them (partial aggregate and partial semi-join push-down) are particularly advantageous in scale-out scenarios where exchanges pose a bottleneck. We propose novel extensions to current literature to enable more aggressive partial push-downs, tailored specifically to the

demands of big-data processing. Lastly, we propose peephole optimizations that customize the implementation of stateful operators based on their input parameters. All of these optimizations are integrated into the Spark engine powering Azure Synapse. Through evaluations conducted using TPCDS, we demonstrate that our optimizations render our engine 1.8 times faster than Apache Spark 3.0.1.

5.

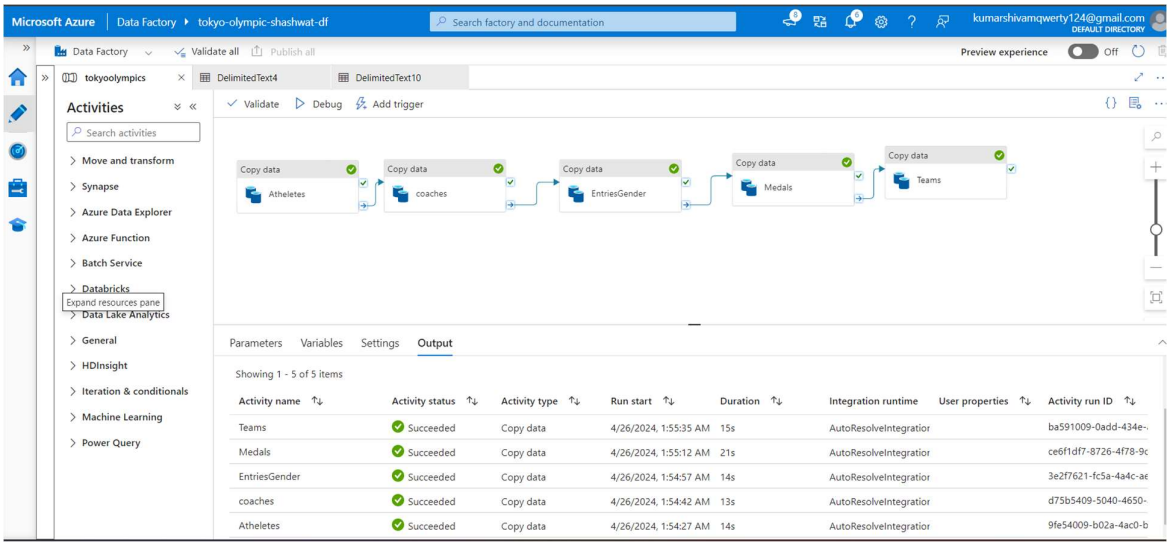
TOPIC: Memristor synapse-coupled piecewise-linear simplified Hopfield neural network: Dynamics analysis and circuit implementation

ABSTRACT: The interaction between adjacent neurons in a neural network generates electromagnetic induction current, facilitated by differences in membrane potential. The memristor, considered the fourth fundamental electric element, mimics neural synapse behaviour and replicates the electromagnetic induction effect. This paper proposes a simplified implementation of a Hopfield neural network (HNN) by replacing the commonly used hyperbolic tangent activation function with a piecewise-linear function. It introduces a straightforward bi-neuron-based memristor synapse-coupled HNN. Theoretical analysis and numerical simulations demonstrate that the memristive HNN exhibits five equilibria, including one unstable saddle-focus, two unstable saddle points, and two stable node points (or node-foci). Local attraction basins and phase plane plots illustrate the multistability of coexisting chaos, periodic limit cycles, and stable point attractors within the memristive HNN model. Various system dynamical behaviors influenced by parameters such as the piecewise-linear activation function, memristor coupling strength, and initial neuron conditions are examined through numerical simulations. Additionally, a simplified analog circuit of the memristive HNN model is designed, leveraging op-amp-based modules for easier hardware implementation. Experimental results validate the accuracy of the design and analyses.

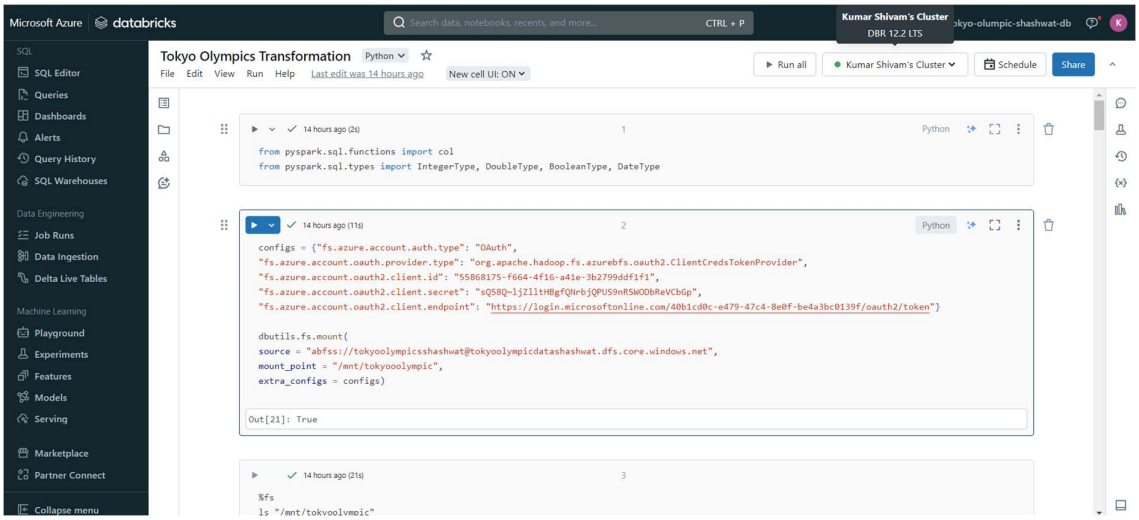
Title	Authors	Year	Techniques and Algorithm Used	Advantages	Disadvantages
Azure Synapse Analytics Use Cases and Reference Architecture	Bhadresh Shiyal	2021	N/A	Provides comprehensive coverage of Azure Synapse Analytics components.	May lack detailed technical implementations.
Hyperspace: the indexing subsystem of azure synapse	Rahul Potharaju, Terry Kim, Eunjin Song, Wentao Wu, Lev Novik, Apoorve Dave, Andrew Fogarty, Pouria Pirzadeh, Vidip Acharya, Gurleen Dhody, Jiying Li	2021	Indexing subsystem design, concurrency control, query processing techniques	Enables building secondary indexes, concurrency control, automatic query acceleration	Requires careful management of indexes, potential overhead in index maintenance
PolyBase in Azure Synapse Analytics	Kevin Feasel	2019	PolyBase functionality for ETL, integration with Azure SQL Data Warehouse	Allows querying across diverse data sources, facilitates ETL processes	May introduce complexity in managing different data sources, potential performance overheads
New query optimization techniques in the Spark engine of Azure synapse	Abhishek Modi, Kaushik Rajan, Srinivas Thimmaiah, Prakhar Jain, Swinky Mann, Ayushi Agarwal, Ajith Shetty, Shahid K I, Ashit Gosalia, Partho Sarthi	2021	Query optimization techniques, exchange placement, partial push-down optimizations, peephole optimizations	Significantly improves query execution speed, specialized optimizations for big data processing	Requires careful tuning, may introduce complexity in Spark engine
Memristor synapse-coupled piecewise-linear simplified Hopfield neural network: Dynamics analysis and circuit implementation	Shoukui Ding, Ning Wang, Han Bao, Bei Chen, Huagan Wu, Quan Xu	2022	Memristor synapse-coupled Hopfield neural network with piecewise-linear activation function	Simplifies implementation, exhibits multistability of chaos and stable attractors	Limited to analog circuit implementation, may require further optimization for practical applications

METHODOLOGY AND OUTPUTS

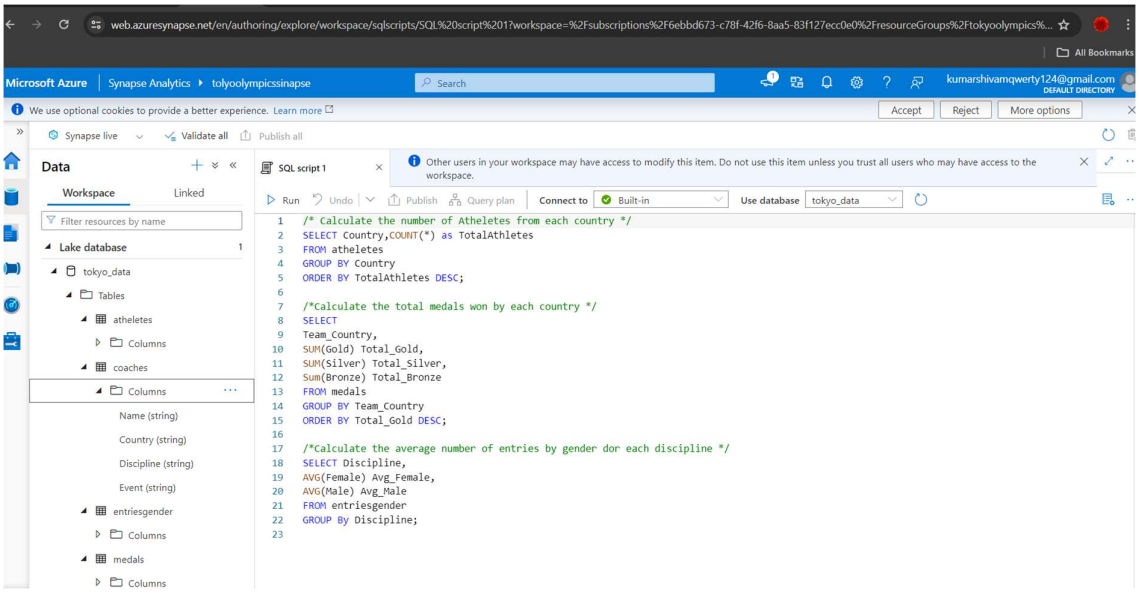
Step 1: Ingesting the data from the Github using Data Factory and putting in the storage accounts using Data Lake .



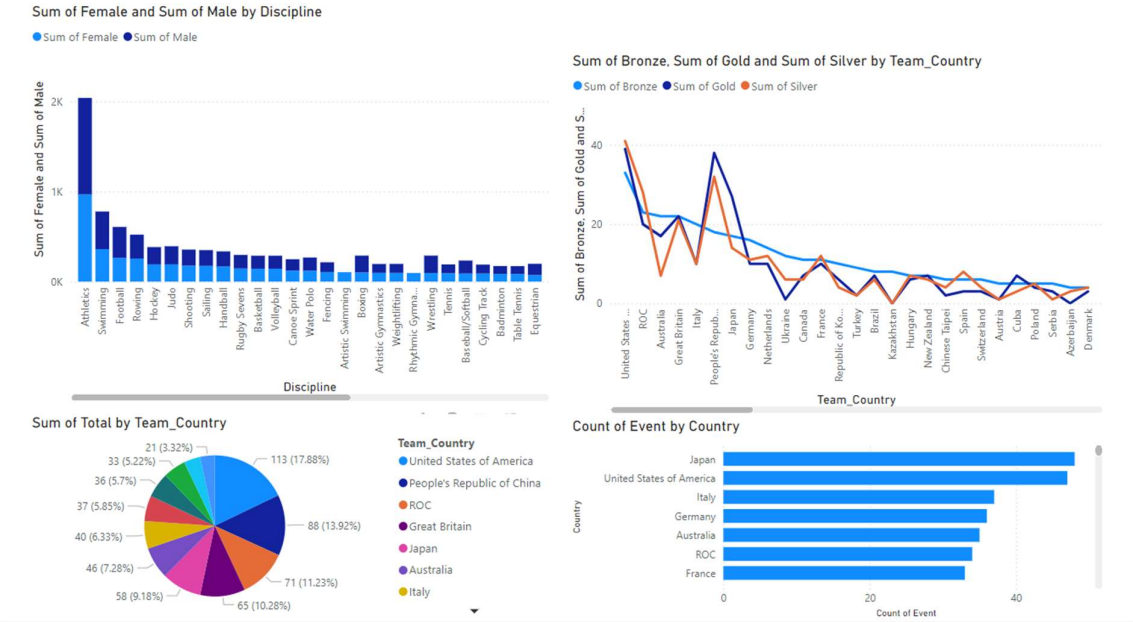
Step 2: Transforming the data using Data Bricks and putting it in Data Lake as Transformed Data.



Step 3: Fetching the data int the Azure Synapse Analysis and doing the Analysis.



Step 4: Linking the Power BI with the Synapse and creating the visualisation dashboard



REFERENCES

1. https://link.springer.com/chapter/10.1007/978-1-4842-7061-5_10

2. https://dl.acm.org/doi/abs/10.14778/3476311.3476382

3. https://link.springer.com/chapter/10.1007/978-1-4842-5461-5_9

4. https://dl.acm.org/doi/abs/10.14778/3503585.3503601

5. https://www.sciencedirect.com/science/article/abs/pii/S0960077922010785