**“Configuration of Azure Data factory and create pipelines to take data from Azure Blob and insert it into Azure SQL”**

*An*

***Internship Project Report***

*submitted in partial fulfillment of the*

*requirements for the internship*

*at*

**Celebal Technologies**

By

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Roll Number** | **Sap ID** |
| Shivam Chaudhary | R970218049 | 500068546 |
| Siddhi Garg | R114218022 | 500069229 |
| Vibhav Sharma | R171218111 | 500068472 |

*under the guidance of*

**Dr. Kingshuk Srivastava**

upes-new-logo

**Department of Informatics**

**School of Computer Science**

**University of Petroleum & Energy Studies**

**Bidholi, Via Prem Nagar, Dehradun, UK**

**DECLARATION**

I/We hereby certify that the project work entitled **“Configuration of Azure Data factory and create pipelines to take data from Azure Blob and insert it into Azure SQL”** in partial fulfillment of the requirements for the internship at Celebal technologies required in completion of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING. School of Computer Science, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my/ our work carried out during a period from **June 7, 2021** to **21July, 2021** under the supervision of **Dr. Kingshuk Srivastava.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Shivam Chaudhary** | **Siddhi Garg** | | **Vaibhav Sharma** |  |
|  | **500068546** | **500069229** | **500068472** | |  |

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

**Date: 04**thAugust, 2021 **Dr. Kingshuk Srivastava**

Project Mentor

**Mr. Tushar Mittal**

Team lead At

Celebal Technologies

upes-new-logo

**ACKNOWLEDGEMENT**

We wish to express our deep gratitude to our mentor **Dr. Kingshuk Srivastava** for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thank to our respected Head of Department, Informatics, **Dr. Thipendra Pal Singh,** for his guidance and support as and when required.

We are also grateful to **Dr. Manish Prateek, Dean SCS**, for providing the necessary facilities to carry out our project work successfully.

We would like to thank all our friends for their help and constructive criticism during our project work. Finally, we have no words to express our sincere gratitude to our **parents** who have shown us this world and for everything they have given to us.

|  |  |  |
| --- | --- | --- |
| **Shivam Chaudhary** | **Siddhi Garg** | **Vaibhav Sharma** |
| **500068546** | **500069229** | **500068472** |

**ABSTRACT**

In this project Configuration of Azure Data factory and create pipelines to take data from Azure Blob and insert it into Azure SQL, create a data factory by using the Azure Data Factory user interface (UI). The pipeline in this data factory copies data from Azure Blob storage to a database in Azure SQL Database. The configuration pattern in this interface applies to copying from a file-based data store to a relational data store. For a list of data stores supported as sources and sinks, and supported data storing table.

It is an automation pipeline system which is used to store the data from one storage system to another storage system.

**INTRODUCTION**

The project is “Configuration of Azure Data factory and create pipelines to take data from Azure Blob and insert it into Azure SQL”.

Azure Data Factory is a managed cloud service that's built for these complex hybrid extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects.

Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers.

Enterprises have data of various types that are located in disparate sources on-premises, in the cloud, structured, unstructured, and semi-structured, all arriving at different intervals and speeds.

The first step in building an information production system is to connect to all the required sources of data and processing, such as software-as-a-service (SaaS) services, databases, file shares, and FTP web services. The next step is to move the data as needed to a centralized location for subsequent processing.

Data Factory offers full support for CI/CD of your data pipelines using Azure DevOps and GitHub. This allows you to incrementally develop and deliver your ETL processes before publishing the finished product.

An Azure subscription might have one or more Azure Data Factory instances (or data factories). Azure Data Factory is composed of below key components.

* Pipelines
* Activities
* Datasets
* Linked services
* Data Flows
* Integration Runtimes

These components work together to provide the platform on which you can compose data-driven workflows with steps to move and transform data.

Blob storage is a feature in Microsoft Azure that lets developers store unstructured data in Microsoft's cloud platform. This data can be accessed from anywhere in the world and can include audio, video and text. Blobs are grouped into "containers" that are tied to user accounts

Azure SQL Database is a fully managed platform as a service (PaaS) database engine that handles most of the database management functions such as upgrading, patching, backups, and monitoring without user involvement.

We need to create a pipeline which copies data from azure blob to azure SQL

**OBJECTIVE**

* Creating a Data Factory pipeline that copies data from Azure Blob Storage to Azure SQL Database
* Azure SQL Database is one of the most popular data repositories for hosting structured data on the Azure cloud.
* Azure Data Factory on Azure cloud helps to transport and transform data from-and-go to a variety of data repositories.
* The Copy Activity performs the data movement in Azure Data Factory. It is powered by a globally available service that can copy data between various data stores in a secure, reliable, and scalable way.

**METHODOLOGY**

1. Azure Data Factory is the cloud-based ETL and data integration service that allows creating data-driven workflows for orchestrating data movement and transforming data at scale*.*
2. Using Azure Data Factory, you can create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.
3. We can publish your transformed data to data stores such as Azure Synapse Analytics for business intelligence (BI) applications to consume. Ultimately, through Azure Data Factory, raw data can be organized into meaningful data stores and data lakes for better business decisions.

**CONCLUSION**

* The setup of Azure blob storage and Azure SQL Database as the source and destination.
* We used Azure Data Factory as the integrator to source data from a CSV file hosted in a container on Azure blob storage, configured and created the data pipeline, and populated this data in Azure SQL Database as the destination for this pipeline.
* Used the different artifacts of Azure Data Factory like Linked Services, Pipelines, Azure Data Factory instance and configuration settings

**REFERENCES**

1. [**https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-database-paas-overview**](https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-database-paas-overview)
2. [**https://docs.microsoft.com/en-us/azure/data-factory/introduction**](https://docs.microsoft.com/en-us/azure/data-factory/introduction)
3. [**https://docs.microsoft.com/en-us/azure/data-factory/tutorial-copy-data-portal**](https://docs.microsoft.com/en-us/azure/data-factory/tutorial-copy-data-portal)
4. **[https://www.sqlshack.com/populate-azure-sql-database-from-azure-blob-storage-using-azure-data-factory/](https://www.sqlshack.com/populate-azure-sql-database-from-azure-blob-storage-using-azure-data-factory/%0c)**