Microsoft Azure Data Factory

Sowmya Ravi[1]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *Corresponding authors: laszewski@gmail.com*

project-000, March 5, 2017

Azure Data Factory is a cloud based data integration service that handles and automates data movement and data integration services. The Data Factory stands out in its ability to ingest data from multiple sources, efficiently build pipelines to manage processes and visualize the results in a pragmatic manner. This paper discusses the data movement and data transformation activities supported by Azure Data Factory

Keywords: Cloud, Data, Movement, Transformaion

https://github.com/sowravi

This review document is provided for you to achieve your best. We have listed a number of obvious opportunities for improvement. When improving it, please keep this copy untouched and instead focus on improving report.tex. The review does not include all possible improvement suggestions and if you sea comment you may want to check if this comment applies elsewhere in the document.

1. INTRODUCTION

The cloud based data integration service is designed to pull data from various sources, transform/ process data and publish the result to the data stores. A data management gateway enables access to data on SQL Databases [**?**]. The data processing is done by It works by creating pipelines to transform the raw data into a format that can be readily used by BI Tools or applications. The services comes with rich visualization aids that aid data analysis. Data Factory supports two types of activities: data movement and data transformation. Data Movement [**?**] is a Copy Activity in Data Factory that copies data from a data source to a Data sink. Data Factory supports the following data stores. Data from any source can be written to any sink. Data Transformation: Azure Data Factory supports the following transformation activities such as Map reduce, Hive transformations and Machine learning activities. Data factory is a great tool to analyze web data, sensor data and geo-spatial data.

A small description for Fig. 1 would help the explanation

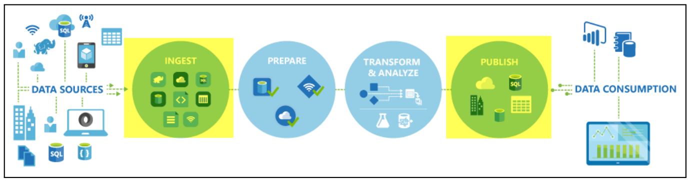


Fig. 1. Work flow in Azure Data Factory [**?**]

2. DATASETS IN AZURE DATA FACTORY

The JSON format is used to store data in Azure Data Factory [**?**].In Azure data factory creating a dataset essentially means creating a pointer to the data to be processed. The data store could be tables, files, folders, and documents. After a dataset is created, it can be used with activities in a pipeline which will be discussed later in the paper. For example, a dataset can be an input/output dataset of a Copy Activity or an HDInsightHive Activity. A visual layout of all your pipelines and data inputs and outputs is also provided by Azure. It enables one all the relationships and dependencies of pipelines across all sources. Thus, it is possible to keep track of all data sources and sinks.

3. PIPELINES AND ACTIVITIES IN AZURE DATA FACTORY

3.1. Data Pipeline

Azure Data Factory used pipelines to organize work flow and data processing [**?**]. A pipeline is an ordering of processes in accordance with the logic relation existing between the processes.

3.2. Activity

An activity is a process or action that is taken on the data [**?**]. An activity takes zero or more datasets as inputs and produces zero or more datasets as an output. The activities are grouped under data movement of data transformation activities.

3.2.1. Data Movement

The Data Movement is achieved by the Data Factory copy wizard. Data can be copied either from local or cloud sources with ease for further transformation to be carried out. The copy activity is highly secure, reliable and scalable

When the source as well as the sink data stores are in the cloud, Copy Activity goes through the following stages to copy data.

- Reads data from the source data store
- Performs serialization/de-serialization, compression/decompression, column mapping, and type conversion. It does these operations based on the configurations of the input dataset, output dataset, and Copy Activity.
- Writes data to the destination data store

In case of moving a data located on-premises to cloud, data management gateway must be installed in the local machine containing the data.The data management gateway is required to perform e serialization/de-serialization, compression/decompression, column mapping, and type conversion. The data management gateway directly writes the data to the destination. The following figure shows the copy activity from a local machine

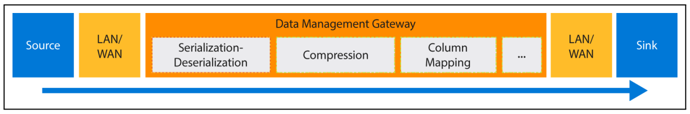


Fig. 2. Moving data between on-premises and cloud data stores [?]

The Azure Data Factory is connected to a variety of data stores and supports different data formats [?]. Copy Activity also reads from and writes to files in specified formats: text, Avro, ORC, Parquet, and JSON, and compression codec GZip, Deflate, BZip2, and ZipDeflate are supported. Furthermore, [?] pipelines could be created either using the copy wizard or using the JSON scripts. The scheduling facility make it possible to run multiple copy operations one after the other in a sequential manner. After copying, the data is ready for transformation purpose

3.2.2. Data Transformation

Data transformation activities transform the raw data to derive predictions and insights. The Data Factory comes with two types of computing environments

- On Demand: This type of environment is solely managed by Data Factory. The environment is created before a process begins and is automatically removed when the process is completed
- : Bring your own: An individual can register his/her computing environment and has to be managed by the individual.

The following transformations are supported By Azure Data Factory [?]

- HDInsight Hive activity: The HDInsight Hive activity executes Hive queries on your own or on-demand Windows/Linux-based HDInsight cluster [?].
- HDInsight Pig activity: The HDInsight Pig activity in a Data Factory pipeline executes Pig queries on your own or on-demand Windows/Linux-based HDInsight cluster [?].
- HDInsight MapReduce activity: The HDInsight MapReduce activity in a Data Factory pipeline executes MapReduce programs on your own or on-demand Windows/Linux-based HDInsight cluster [?].
- HDInsight Streaming activity: HDInsightStreamingActivity Activity is used to invoke a Hadoop Streaming job from an Azure Data Factory pipeline. The HDInsight Streaming Activity in a Data Factory pipeline executes Hadoop Streaming programs on your own or on-demand Windows/Linux-based HDInsight cluster [?].
- Machine Learning activities: Azure Data Factory helps in building and deploying predictive models in the following three steps [?].
  1. Create a training experiment: This is done using the Azure ML Studio. The ML studio is a collaborative visual development environment that can be used to train and test a predictive analytics model using training data.
  2. Convert it to a predictive experiment: Once the model has been trained with existing data, it can be used to score new data, prepare and streamline the experiment for scoring.
  3. Deploy it as a web service. The experiment can be published using an Azure web service.
- Stored procedure activity: A procedure that is already stored can be invoked in one of the following data stores: Azure SQL Database, Azure SQL Data Warehouse, SQL Server Database in your enterprise or on an Azure virtual machine (VM) [?] .
- Data Lake Analytics U-SQL activity: Azure Data Lake Analytics can be linked to the compute service in the Data Factory [?].
- .NET custom activity: One can create an activity with a customized logic to be used in a pipeline [?].

4. MONITORING AND MANAGEMENT

The Data Factory has a monitor and manage app which helps keep tab of available resources, handle alerts and monitor various views. The resource explorer has a tree view and a diagram view of all the existing pipelines, datasets and linked services [?].

5. USE CASES

Data factory can be used for a wide range of applications such as Product recommendations, Customer profiling etc. The sample solution for a recommendation system is illustrated in Fig.3. In this case, the customer information, product information and the behavioural data of the customer are stored in azure blobs. The combined data is fed to the recommendation system which is modelled using any of the algorithms available in the Azure Machine Learning suite. The final result contains customer specific product recommendations that can be written to relational database for the consumption by the store's website [?].

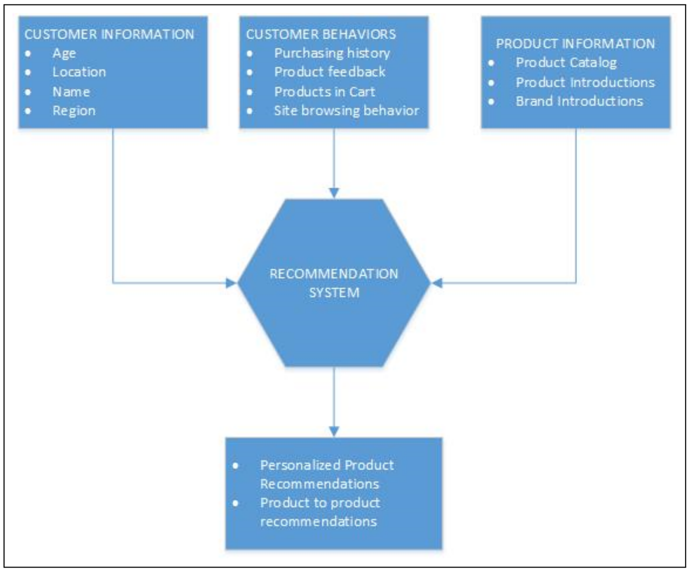
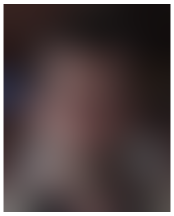


Fig. 3. A recommendation engine using Azure Data factory [?]

6. CONCLUSION

The paper talks about the data orchestration and management using Microsoft's Data Factory. It's ability to link diverse data stores makes it a versatile tool for data integration. In addition, the powerful pipeline based data transformation services embedded in it helps data scientists and developers to provide efficient and effective solutions using large datasets.

AUTHOR BIOGRAPHIES



Sowmya Ravi pursuing Master's in Data Science at Indiana University, Bloomington. Her research interests include Machine learning, Data Mining and Big Data