Microsoft Azure

Data Engineering in Azure [DP-203]

1

About this course

In this course, the student will learn about the data engineering patterns and practices as it pertains to working with batch and real-time analytical solutions using Azure data platform technologies. Students will begin by understanding the core compute and storage technologies that are used to build an analytical solution. They will then explore how to design an analytical serving layers and focus on data engineering considerations for working with source files.

The students will learn how to interactively explore data stored in files in a data lake. They will learn the various ingestion techniques that can be used to load data using the Apache Spark capability found in Azure Synapse Analytics or Azure Databricks, or how to ingest using Azure Data Factory or Azure Synapse pipelines. The students will also learn the various ways they can transform the data using the same technologies that is used to ingest data.

The student will spend time on the course learning how to monitor and analyze the performance of analytical system so that they can optimize the performance of data loads, or queries that are issued against the systems. They will understand the importance of implementing security to ensure that the data is protected at rest or in transit.

The student will then show how the data in an analytical system can be used to create dashboards or build predictive models in Azure Synapse Analytics.

2

© Microsoft Corporation

Course agenda

Module 0
Data Engineering in Azure course overview

Lesson 01 – About this course

Lesson 02 – Course agenda

Lesson 03 – Audience

Lesson 04 – Prerequisites

Lesson 05 – Course lab setup

Module 1
Explore compute and storage options for data engineering workloads

Lesson 01 – Introduction to Azure Synapse Analytics

Lesson 02 – Describe Azure Databricks

Lesson 03 – Introduction to Azure Data Lake storage

Lesson 04 – Describe Delta Lake architecture

Lesson 05 – Work with data streams by using Azure Stream Analytics

3

Course agenda *(continued #1)*

Module 2
Run interactive queries using Azure Synapse Analytics serverless SQL pool

Lesson 01 – Explore Azure Synapse serverless SQL pools capabilities

Lesson 02 – Query data in the lake using Azure Synapse serverless SQL pools

Lesson 03 – Create metadata objects in Azure Synapse serverless SQL pools

Lesson 04 – Secure data and manage users in Azure Synapse serverless SQL pools

Module 3
Data exploration and transformation in Azure Databricks

Lesson 01 – Describe Azure Databricks

Lesson 02 – Read and write data in Azure Databricks

Lesson 03 – Work with DataFrames in Azure Databricks

Lesson 04 – Work with DataFrames advanced methods in Azure Databricks

4

© Microsoft Corporation

Course agenda *(continued #2)*

Module 4
Explore, transform, and load data into the data warehouse using Azure Synapse Analytics Apache Spark

Lesson 01 – Understand big data engineering with Apache Spark in Azure Synapse Analytics

Lesson 02 – Ingest data with Apache Spark notebooks in Azure Synapse Analytics

Lesson 03 – Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics

Lesson 04 – Integrate SQL and Apache Spark pools in Azure Synapse Analytics

Module 5
Ingest and load data into the data warehouse

Lesson 01 – Use data loading best practices in Azure Synapse Analytics

Lesson 02 – Petabyte-scale ingestion with Azure Data Factory or Azure Synapse Pipelines

5

Course agenda *(continued #3)*

Module 6
Transform data with Azure Data Factory or Azure Synapse Pipelines

Lesson 01 – Data integration with Azure Data Factory or Azure Synapse Pipelines

Lesson 02 – Code-free transformation at scale with Azure Data Factory or Azure Synapse Pipelines

Module 7
Orchestrate data movement and transformation in Azure Data Factory or Azure Synapse Pipelines

Lesson 01 – Orchestrate data movement and transformation in Azure Data Factory or Azure Synapse Pipelines

6

© Microsoft Corporation

Course agenda *(continued #4)*

Module 8
End-to-end security with Azure Synapse Analytics

Lesson 01 – Secure a data warehouse in Azure Synapse Analytics

Lesson 02 – Configure and manage secrets in Azure Key Vault

Lesson 03 – Implement compliance controls for sensitive data

Module 9
Support Hybrid Transactional Analytical Processing (HTAP) with Azure Synapse Link

Lesson 01 – Design hybrid transactional and analytical processing using Azure Synapse Analytics

Lesson 02 – Configure Azure Synapse Link with Azure Cosmos DB

Lesson 03 – Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics

Lesson 04 – Query Azure Cosmos DB with SQL serverless for Azure Synapse Analytics

7

Course agenda *(continued #5)*

Module 10
Real-time stream processing with Azure Stream Analytics

Lesson 01 – Enable reliable messaging for Big Data applications using Azure Event Hubs

Lesson 02 – Work with data streams by using Azure Stream Analytics

Lesson 03 – Ingest data streams with Azure Stream Analytics

Module 11
Create a stream processing solution with Event Hubs and Azure Databricks

Lesson 01 – Process streaming data with Azure Databricks structured streaming

8

© Microsoft Corporation

Audience

Primary audience:
The audience for this course are data engineers, data professionals, data architects, and business intelligence professionals who want to learn about the data platform technologies that exist on Microsoft Azure that can be used to perform data engineering and storage for analytical solutions.

Secondary audience:
The secondary audience for this course are individuals who develop applications that deliver content from the data platform technologies that exist on Microsoft Azure.

9

Microsoft Azure

Module 01 : Explore compute and storage options for data engineering workloads

10

© Microsoft Corporation

Lesson 01: Introduction to Azure Synapse Analytics

11

Azure Synapse Analytics

Limitless analytics service with unmatched time to insight

12

© Microsoft Corporation

Introduction to Azure Synapse Analytics

13

Modern data warehousing pattern with Azure Synapse Analytics

14

© Microsoft Corporation

Lesson 02: Describe Azure Databricks

15

What is Azure Databricks

Apache Spark-based analytics platform:
Simplifies the provisioning and collaboration of Apache Spark-based analytical solutions dealing with batch and streaming data

Comprehensive Spark library support:
Support includes SQL, DataFrames, MLlib, Hyperspace and MSSparkUtil

Enterprise Security:
Utilizes the security capabilities of Azure

Integration with other Cloud Services:
Can integrate with a variety of Azure data platform services and Power BI

16

© Microsoft Corporation

What is Apache Spark

Apache Spark emerged to provide a parallel processing framework that supports in-memory processing to boost the performance of big-data analytical applications on massive volumes of data

Interactive Data Analysis:
Used by business analysts or data engineers to analyze and prepare data

Streaming Analytics:
Ingest data from technologies such as Kafka and Flume to ingest data in real-time

Machine Learning:
Contains a number of libraries that enables a Data Scientist to perform Machine Learning

Why use Azure Databricks?
Azure Databricks is a wrapper around Apache Spark that simplifies the provisioning and configuration of a Spark cluster in a GUI interface

Azure Databricks components:
Spark SQL and DataFrames
Streaming
Mlib
GraphX
Spark Core API

17

Enterprise security

18

© Microsoft Corporation

Integration with cloud services

19

Spark: what to use when and where

	Apache Spark	HDInsight	Azure Databricks	Synapse Spark
WHAT	Is an Open Source memory optimized system for managing big data workloads	Microsoft implementation of Open Source Spark managed within the realms of Azure	A managed Spark as a Service solution	Embedded Spark capability within Azure Synapse Analytics
WHEN	When you want to benefits of spark for big data processing and/or data science work without the Service Level Agreements of a provider	When you want to benefits of OSS spark with the Service Level Agreement of a provider	Provides end to end data engineering and data science solution and management platform	Enables organizations without existing Spark implementations to fire up a Spark cluster to meet data engineering needs without the overheads of the other Spark platforms listed
WHO	Open Source Professionals	Open Source Professionals wanting SLA's and Microsoft Data Platform experts	Data Engineers and Data Scientists working on big data projects every day	Data Engineers, Data Scientists, Data Platform experts and Data Analysts
WHY	To overcome the limitations of SMP systems imposed on big data workloads	To take advantage of the OSS Big Data Analytics platform with SLA's in place to ensure business continuity	It provides the ability to create and manage an end to end big data/data science project using one platform	It provides the ability to scale efficiently with spark clusters within a one stop shop Data Warehousing platform of Synapse.

20

© Microsoft Corporation

Lesson 03: Introduction to Azure Data Lake storage

21

Processing Big Data with Azure Data Lake Store

22

© Microsoft Corporation

Introduction to Azure Data Lake storage

- Hadoop compatible
- Security
- Performance
- Redundancy

23

Compare Azure Blob Storage and Data Lake Store Gen 2

Azure Blob Flat namespace	Data Lake (Gen II) Hierarchical namespace

24

© Microsoft Corporation

Big Data use cases

Let's examine three use cases for leveraging an Azure Data Lake Store

Modern data warehouse	This architecture sees Azure Data Lake Storage at the heart of the solution for a modern data warehouse. Using Azure Data Factory to ingest data into the Data Lake from a business application, and predictive models built in Azure Databricks, using Azure Synapse Analytics as a serving layer
Advanced analytics	In this solution, Azure Data factory is transferring terabytes of web logs from a web server to the Data Lake on an hourly basis. This data is provided as features to the predictive model in Azure Databricks, which is then trained and scored. The result of the model is then distributed globally using Azure Cosmos DB, that an application uses
Real time analytics	In this architecture, there are two ingestion streams. Azure Data Factory is used to ingest the summary files that are generated when the HGV engine is turned off. Apache Kafka provides the real-time ingestion engine for the telemetry data. Both data streams are stored in Data Lake store for use in the future

25

Lesson 04: Describe Delta Lake architecture

26

© Microsoft Corporation

Describe a Delta Lake architecture

27

Lesson 05: Work with data streams by using Azure Stream Analytics

28

© Microsoft Corporation

What are data streams

Data streams:
In the context of analytics, data streams are event data generated by sensors or other sources that can be analyzed by another technology

Data stream processing approach:
There are two approaches. Reference data is streaming data that can be collected over time and persisted in storage as static data. In contrast, streaming data have relatively low storage requirements. And run computations in sliding windows

Data streams are used to:

Analyze data:	Understand systems:	Trigger actions:
Continuously analyze data to detect issues and understand or respond to them	Understand component or system behavior under various conditions to fuel further enhancements of said system	Trigger specific actions when certain thresholds are identified

29

Event processing

The process of consuming data streams, analyzing them, and deriving actionable insights out of them is called Event Processing and has three distinct components:

Event producer	Examples include sensors or processes that generate data continuously such as a heart rate monitor or a highway toll lane sensor
Event processor	An engine to consume event data streams and deriving insights from them. Depending on the problem space, event processors either process one incoming event at a time (such as a heart rate monitor) or process multiple events at a time (such as a highway toll lane sensor)
Event consumer	An application which consumes the data and takes specific action based on the insights. Examples of event consumers include alert generation, dashboards, or even sending data to another event processing engine

30

© Microsoft Corporation

Processing events with Azure Stream Analytics

Microsoft Azure Stream Analytics is an event processing engine. It enables the consumption and analysis of high volumes of streaming data in real time

Source	Ingestion	Analytical engine	Destination
Sensors	Event Hubs	Stream Analytics Query Language	Azure Data Lake
Systems	IoT Hubs	.NET SDK	Cosmos DB
Applications	Azure Blob Store		SQL Database
			Blob Store
			Power BI

31

Work with data streams by using Azure Stream Analytics

Complex event processing of Stream Data in Azure

32

© Microsoft Corporation

Azure Data Platform technologies

Storage Account

- When you need a **low cost, high throughput** data store
- When you need to store **No-SQL** data
- When you **do not need to query** the data directly. **No ad hoc query** support
- Suits the storage of archive or **relatively static data**
- Suits acting as a **HDInsight Hadoop** data store

Data Lake Store

- When you need a **low cost, high throughput** data store
- **Unlimited storage for No-SQL** data
- When you **do not need to query** the data directly. **No ad hoc query** support
- Suits the storage of archive or **relatively static data**
- Suits acting as a **Databricks , HDInsight** and **IoT** data store

Azure Databricks

- **Eases the deployment** of a Spark based cluster
- Enables the **fastest processing** of Machine Learning solutions
- **Enables collaboration** between data engineers and data scientists
- Provides **tight enterprise security integration** with Azure Active Directory
- **Integration with other Azure Services and Power BI**

Azure COSMOS DB

- Provides **global distribution** for both structured and unstructured data stores
- **Millisecond query response** time
- **99.999% availability** of data
- **Worldwide elastic scale** of both the storage and throughput
- **Multiple consistency levels** to control data integrity with concurrency

Azure SQL Database

- When you require a **relational** data store
- When you need to manage **transactional workloads**
- When you need to manage a **high volume on inserts and reads**
- When you need a service that **requires high concurrency**
- When you require a solution that can scale **elastically**

33

Azure Data Platform technologies (continued)

Azure Synapse Analytics

- When you require an integrated **relational** and **big data** store
- When you need to manage **data warehouse** and **analytical workloads**
- When you need **low cost storage**
- When you require the ability to **pause and restart the compute**
- When you require a solution that can scale **elastically**

Azure Stream Analytics

- When you require a **fully managed event processing** engine
- When you require **temporal analysis of streaming** data
- Support for analyzing **IoT streaming** data
- Support for analyzing application data through **Event Hubs**
- Ease of use with a **Stream Analytics Query Language**

Azure Data Factory

- When you want to **orchestrate the batch movement** of data
- When you want to connect to **wide range of data platforms**
- When you want to **transform or enrich** the data in movement
- When you want to **integrate with SSIS packages**
- Enables **verbose logging** of data processing activities

Azure HDInsight

- When you need a **low cost, high throughput** data store
- When you need to store **No-SQL** data
- Provides a Hadoop **Platform as a Service** approach
- Suits acting as a **Hadoop, Hbase, Storm or Kafka** data store
- **Eases the deployment and management** of clusters

Azure Purview

- When you require **documentation** of your data stores
- When you require a **multi user** approach to documentation
- When you need to **classify data sources**
- A **fully managed cloud service** whose users can discover the data sources
- When you require **a solution that can help any user** understand their data

34

© Microsoft Corporation

Lab: Explore compute and storage options for data engineering workloads

35

Lab overview

This lab teaches ways to structure the data lake, and to optimize the files for exploration, streaming, and batch workloads. The student will learn how to organize the data lake into levels of data refinement as they transform files through batch and stream processing. The students will also experience working with Apache Spark in Azure Synapse Analytics. They will learn how to create indexes on their datasets, such as CSV, JSON, and Parquet files, and use them for potential query and workload acceleration using Spark libraries including Hyperspace and MSSParkUtils.

Lab objectives

After completing this lab, you will be able to:

- Work with a Delta Lake architecture
- Working with Apache Spark in Azure Synapse Analytics

36

© Microsoft Corporation

37

Microsoft Azure

Module 02 : Run interactive queries using Azure Synapse Analytics serverless SQL pools

38

© Microsoft Corporation

Lesson 01: Explore Azure Synapse serverless SQL pools capabilities

39

Azure Synapse serverless SQL Pools

Every Azure Synapse Analytics workspace comes with serverless SQL pool endpoints so you can start querying data in seconds to minutes in a data lake as soon as the workspace is created. There's no infrastructure to setup or clusters to maintain.

40

© Microsoft Corporation

Comparing dedicated SQL Pools with serverless SQL pools in Azure Synapse Analytics

Dedicated SQL pools

- Used for Data Warehouse operations
- Provides predictable performance and costs
- Reserves processing power for data stored in SQL tables

Serverless SQL pools

- Used for data preparation or ad-hoc queries against unstructured data.
- Provides an always available SQL endpoint for unplanned workloads
- Enables interactive querying

41

Explore Azure Synapse serverless SQL pools capabilities

Every Azure Synapse Analytics workspace comes with serverless SQL pool endpoints so you can start querying data in seconds to minutes in a data lake as soon as the workspace is created. There's no infrastructure to setup or clusters to maintain.

Data Exploration	Data transformation	Logical data warehouse
Browse the data lake and get initial insights about the data. Using Azure Synapse Studio, you can explore the data both graphically and programmatically.	Serverless SQL pool enables you to execute transformation statements over the data in the lake and store the results back to the data lake in a specified file format.	Create objects (such as VIEWs and External Tables) that provide you with a SQL metadata layer over the data in the lake to create a logical data warehouse. Once these objects are created, any tool that can connect to serverless SQL pool will see these objects as regular SQL Server objects

42

© Microsoft Corporation

Lesson 02: Query data in the lake using Azure Synapse serverless SQL pools

43

Common files to query

Parquet

Json

DelimitedText

44

© Microsoft Corporation

Using Azure Synapse Studio to view data

45

Querying parquet files in a data lake

46

© Microsoft Corporation

Lesson 03: Create metadata objects in Azure Synapse serverless SQL pools

47

Create metadata objects in Azure Synapse serverless SQL pools

Database

```
CREATE DATABASE [YourDatabaseName]
```

Database scoped credential

```
CREATE DATABASE SCOPED CREDENTIAL [sqlondemand]
WITH IDENTITY='SHARED ACCESS SIGNATURE',
SECRET = 'sv=2018-03-28&ss=bf&srt=sco&sp=rl&'
```

External data source

```
CREATE EXTERNAL DATA SOURCE SqlOnDemandDemo WITH (
   LOCATION = 'https://sqlondemandstorage.blob.core.windows.net',
   CREDENTIAL = sqlondemand );
```

External file format

```
CREATE EXTERNAL FILE FORMAT QuotedCsvWithHeaderFormat
WITH
( FORMAT_TYPE = DELIMITEDTEXT,
FORMAT_OPTIONS ( FIELD_TERMINATOR = ',', STRING_DELIMITER = '"',
FIRST_ROW = 2 ) );
```

External Table

```
CREATE EXTERNAL TABLE populationExternalTable
( [country_name] VARCHAR (100) COLLATE Latin1_General_BIN2,
[year] smallint, [population] bigint )
WITH
( LOCATION = 'csv/population/population.csv',
DATA_SOURCE = sqlondemanddemo,
FILE_FORMAT = QuotedCSVWithHeaderFormat );
```

48

© Microsoft Corporation

Lesson 04: Secure data and manage users in Azure Synapse serverless SQL pools

49

Securing access to data in a data lake when using Azure Synapse Analytics

50

© Microsoft Corporation

Choose an authentication method

51

Manage users in Azure Synapse serverless SQL pools

52

© Microsoft Corporation

Manage user access to data lake files

53

Lab: Run interactive queries using Azure Synapse Analytics serverless SQL pools

54

© Microsoft Corporation

Lab overview

In this lab, students will learn how to work with files stored in the data lake and external file sources, through T-SQL statements executed by a serverless SQL pool in Azure Synapse Analytics. Students will query Parquet files stored in a data lake, as well as CSV files stored in an external data store. Next, they will create Azure Active Directory security groups and enforce access to files in the data lake through Role-Based Access Control (RBAC) and Access Control Lists (ACLs).

Lab objectives

After completing this lab, you will be able to:

- Querying a Data Lake Store using serverless SQL pools in Azure Synapse Analytics
- Securing access to data through using a serverless SQL pool in Azure Synapse Analytics

55

56

© Microsoft Corporation

Microsoft Azure

Module 03 : Data Exploration and Transformation in Azure Databricks

57

Lesson 01: Understand Azure Databricks

58

© Microsoft Corporation

Understand Azure Databricks

59

Lesson 02: Read and write data in Azure Databricks

60

© Microsoft Corporation

Read and write data in Azure Databricks

Multiple format support

Reading data from CSV, PARQUET, JSON and many others

Integrated with several Azure Data Services

Reading and writing from and to Azure Data Lake Storage, Azure Synapse Analytics, etc.

Notebook experience

Reading and writing by simply writing code in a shared notebook experience

61

Read data in Azure Databricks

	TransactionId	CustomerId	ProductId	Quantity	Price	TotalAmount	TransactionDate	ProfitAmount	Hour	Minute	StoreId
1	e067fc11-e07d-4517-bc93-f7dc4b44f35e	3	4581	4	20.84	91.696	20170501	26.048	2	30	7922
2	e067fc11-e07d-4517-bc93-f7dc4b44f35e	3	1365	4	26.52	116.688	20170501	29.436	2	30	7922
3	e067fc11-e07d-4517-bc93-f7dc4b44f35e	3	2641	4	29.71	130.724	20170501	37.4	2	30	7922
4	e067fc11-e07d-4517-bc93-f7dc4b44f35e	3	220	2	27.6	60.72	20170501	15.356	2	30	7922
5	e067fc11-e07d-4517-bc93-f7dc4b44f35e	3	110	3	28.41	93.753	20170501	33	2	30	7922
6	e067fc11-e07d-4517-bc93-f7dc4b44f35e	3	2	1	39.78	43.758	20170501	11.528	2	30	7922
7	cdd2ed88-8aae-4295-884a-ac4d40c3c33c	11	3323	1	30.52	33.572	20170501	10.252	20	43	3573

62

© Microsoft Corporation

Working with Select in Azure Databricks

SQL	DataFrame (Python)
SELECT col_1 FROM myTable	df.select(col("col_1"))
DESCRIBE myTable	df.printSchema()
SELECT * FROM myTable WHERE col_1 > 0	df.filter(col("col_1") > 0)
..GROUP BY col_2	..groupBy(col("col_2"))
..ORDER BY col_2	..orderBy(col("col_2"))
..WHERE year(col_3) > 1990	..filter(year(col("col_3")) > 1990)
SELECT * FROM myTable LIMIT 10	df.limit(10)
display(myTable) (text format)	df.show()
display(myTable) (html format)	display(df)

63

Write data in Azure Databricks

write a data file (Scala)

Test cluster | File | Edit | View: Standard | Permissions | Run All | Clear

Cmd 1

```scala
spark.conf.set("fs.azure.account.auth.type", "OAuth")
spark.conf.set("fs.azure.account.oauth.provider.type", "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")
spark.conf.set("fs.azure.account.oauth2.client.id.[redacted].dfs.core.windows.net", "[redacted]")
spark.conf.set("fs.azure.account.oauth2.client.secret.[redacted].dfs.core.windows.net", "[redacted]")
spark.conf.set("fs.azure.account.oauth2.client.endpoint.[redacted].dfs.core.windows.net", "https://login.microsoftonline.com/[redacted]/oauth2/token")
```

Cmd 2

```scala
val df = spark.read.option("header",true).csv("abfss://wwi-02@[redacted].dfs.core.windows.net/sale-poc/sale-20170501.csv")
```

Cmd 3

```scala
val df_distinct_products = df.select(df("ProductId")).distinct
```

Cmd 4

```scala
display(df_distinct_products.limit(10))
```

Cmd 5

```scala
df.write.option("header",true)
  .csv("abfss://wwi-02@[redacted].dfs.core.windows.net/sale-poc/distinctproductid.csv")
```

Shift+Enter to run

64

© Microsoft Corporation

Lesson 03: Work with DataFrames in Azure Databricks

65

Work with DataFrames in Azure Databricks

- Apache Spark DataFrame API reading data in a single command

```
parquetDir = source + "/wikipedia/pagecounts/staging_parquet_en_only_clean/"

pagecountsEnAllDF = (spark  # Our SparkSession & Entry Point
  .read                     # Our DataFrameReader
  .parquet(parquetDir)      # Returns an instance of DataFrame
)
print(pagecountsEnAllDF)    # Python hack to see the data type
```

66

© Microsoft Corporation

Working with transformations in Azure Databricks

Transformations	Description
Select(...)	The select(...) command enables you to specify the columns to include in a query
drop(...)	The drop(...) command enables you to specify the columns you don't want
distinct(...)	The distinct(...) command returns a distinct set of values in a DataFrame
dropDuplicates(...)	The dropDuplicates(...) command is an alias of the distinct(...) command.
show(...)	The show(..) command is part of the core Spark API and simply prints the results to the console
display(...)	The display(...) command provides more flexibility than show(...) such as downloading results against csv, rendering charts and showing up to 100 rows
limit(...)	The limit(...) command can be used to control the number of records that are returned to a DataFrame

67

Optimize DataFrames in Azure Databricks

- Mix DataFrame operations

```
(pagecountsEnAllDF
  .cache()          # Mark the DataFrame as cached
  .count()          # Materialize the cache
)
```

68

© Microsoft Corporation

Lesson 04: Work with DataFrames advanced methods in Azure Databricks

69

Work with DataFrames advanced methods in Azure Databricks

DateTime manipulation
Enabling different DateTime techniques to use across DataFrames

Aggregate Functions
groupBy() function, sum(), count(), avg(), min(), max() functions

Deduplication of Data
Removing duplicates, by ensuring you only keep 1 record

70

© Microsoft Corporation

Lab: Data Exploration and Transformation in Azure Databricks

71

Lab overview

This lab teaches you how to use various Apache Spark DataFrame methods to explore and transform data in Azure Databricks. You will learn how to perform standard DataFrame methods to explore and transform data. You will also learn how to perform more advanced tasks, such as removing duplicate data, manipulate date/time values, rename columns, and aggregate data.

Lab objectives

After completing this lab, you will be able to:

- Use DataFrames in Azure Databricks to explore and filter data
- Cache a DataFrame for faster subsequent queries
- Remove duplicate data
- Manipulate date/time values
- Remove and rename DataFrame columns
- Aggregate data stored in a DataFrame

72

© Microsoft Corporation

73

Microsoft Azure

Module 04 : Explore, transform, and load data into the Data Warehouse using Apache Spark

74

© Microsoft Corporation

Lesson 01: Understand big data engineering with Apache Spark in Azure Synapse Analytics

75

Introduction to big data engineering with Apache Spark in Azure Synapse Analytics

76

© Microsoft Corporation

How do Apache Spark pools work in Azure Synapse Analytics

77

How to create an Apache Spark pool in Azure Synapse Analytics

78

© Microsoft Corporation

Lesson 02: Ingest data with Apache Spark notebooks in Azure Synapse Analytics

79

Apache Spark notebooks features in Azure Synapse Analytics

Notebooks

- Access through Synapse Studio
- Examples Available through Knowledge Center
- Allows to write multiple languages in one notebook by using %%<Name of language>
- Support for Language Syntax highlight, syntax error, syntax code completion
- Offers temporary tables across languages
- Export results

80

© Microsoft Corporation

Creating a notebook in Azure Synapse Analytics

81

Ingest data with Apache Spark notebooks in Azure Synapse Analytics

- Generating data while executing the command

- Loading data in a single command from a data file

82

© Microsoft Corporation

Lesson 03: Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics

83

Transform data with DataFrames in Apache Spark Pools in Azure Synapse Analytics

84

© Microsoft Corporation

Lesson 04: Integrate SQL and Apache Spark pools in Azure Synapse Analytics

85

Integrate SQL and Apache Spark pools in Azure Synapse Analytics

86

© Microsoft Corporation

Write data from Apache Spark pools to a dedicated SQL pool

wwi-02 | Notebook 1

Run all | Publish | Outline | Attach to SparkPool01 | Language PySpark (Pyt

Not started

```
# Write data from a dataframe to a table

# Create a temporary view for top purchases so we can load from Scala
topPurchases.createOrReplaceTempView("top_purchases")
```

```
%%spark

val df = spark.sqlContext.sql("select * from top_purchases")
df.write.sqlanalytics("SQLPool01.wwi.TopPurchases", Constants.INTERNAL)
```

87

Write data from a dedicated SQL pool to Apache Spark pools

wwi-02 | Notebook 1

Run all | Publish | Outline | Attach to SparkPool01 | Language

Not started

```
# Write data from a table to a view in Spark
```

```
%%spark
val df2 = spark.read.sqlanalytics("SQLPool01.wwi.TopPurchases")
df2.createTempView("top_purchases_sql")
```

88

© Microsoft Corporation

Lab: Explore, transform, and load data into the Data Warehouse using Apache Spark

89

Lab overview

This lab teaches you how to explore data stored in a data lake, transform the data, and load data into a relational data store. You will explore Parquet and JSON files and use techniques to query and transform JSON files with hierarchical structures. Then you will use Apache Spark to load data into the data warehouse and join Parquet data in the data lake with data in the dedicated SQL pool.

Lab objectives

After completing this lab, you will be able to:

- Perform Data Exploration in Synapse Studio
- Ingest data with Spark notebooks in Azure Synapse Analytics
- Transform data with DataFrames in Spark pools in Azure Synapse Analytics
- Integrate SQL and Spark pools in Azure Synapse Analytics

90

© Microsoft Corporation

91

Microsoft Azure

Module 05 : Ingest and load data into the data warehouse

92

© Microsoft Corporation

Lesson 01: Use data loading best practices in Azure Synapse Analytics

93

Dedicated SQL Pool architecture revision

94

© Microsoft Corporation

Understand data load design goals

- Where is my data coming from?
- Is the data nett new? or do you receive changes from existing datasets?
- How often is the data being refreshed, added to or replaced?
- What formats are the data coming in?
- Is the data ingestible as-is? or are transformations and cleansing tasks required?
- Which takes priority, loading or querying/analysis?

95

Manage singleton updates

96

© Microsoft Corporation

Set-up dedicated data loading accounts

97

Manage concurrent access to Azure Synapse Analytics

98

© Microsoft Corporation

Implement Workload Management

99

Use PolyBase, the Copy command or the Copy Activity

PolyBase

Copy command

Copy data activity

100

© Microsoft Corporation

Lesson 02: Petabyte-scale ingestion with Azure Data Factory

101

Azure Data Factory/Synapse pipeline revision

102

© Microsoft Corporation

Petabyte-scale ingestion with Azure Data Factory

103

Understanding integration

104

© Microsoft Corporation

Lab: Ingest and load data into the data warehouse

105

Lab overview

This lab teaches students how to ingest data into the data warehouse through T-SQL scripts and Synapse Analytics integration pipelines. The student will learn how to load data into Synapse dedicated SQL pools with PolyBase and COPY using T-SQL. The student will also learn how to use workload management along with a Copy activity in a Azure Synapse pipeline for petabyte-scale data ingestion.

Lab objectives

After completing this lab, you will be able to:

- Use data loading best practices in Azure Synapse Analytics
- Petabyte-scale ingestion with Azure Data Factory

106

© Microsoft Corporation

107

Microsoft Azure

Module 06 : Transform data with Azure Data Factory or Azure Synapse Pipelines

108

© Microsoft Corporation

Lesson 01: Data integration with Azure Data Factory or Azure Synapse Pipelines

109

Data integration with Azure Data Factory or Azure Synapse Pipelines

110

© Microsoft Corporation

Linked Services

111

Datasets

112

© Microsoft Corporation

Activities and pipelines

113

Lesson 02: Code-free transformation at scale with Azure Data Factory or Azure Synapse Pipelines

114

© Microsoft Corporation

Code-free transformation at scale with Azure Data Factory or Azure Synapse Pipelines

115

Adding a source

116

© Microsoft Corporation

Schema modifier transformations

117

Formatter transformations

118

© Microsoft Corporation

Multiple inputs/outputs transformations

119

Row modifier transformations

120

© Microsoft Corporation

Sink transformation

121

Lab: Transform data with Azure Data Factory or Azure Synapse Pipelines

122

© Microsoft Corporation

Lab overview

This lab teaches students how to build data integration pipelines to ingest data from multiple data sources, transform data using mapping data flows and notebooks, and perform data movement into one or more data sinks.

Lab objectives

After completing this lab, you will be able to:

- Data integration with Azure Data Factory or Azure Synapse Pipelines
- Code-free transformation at scale with Azure Data Factory or Azure Synapse Pipelines

123

124

© Microsoft Corporation

Microsoft Azure

Module 07 : Integrate data from Notebooks with Azure Data Factory or Azure Synapse Pipelines

125

Lesson 01: Integrate data from Notebooks with Azure Data Factory or Azure Synapse Pipelines

126

© Microsoft Corporation

Integrating data from Notebooks with Azure Data Factory or Azure Synapse Pipelines

127

Programmatically creating Azure Data Factory or Azure Synapse Pipelines

128

© Microsoft Corporation

Create a notebook

Cell 1

```
%%pyspark
df = spark.read.load('abfss://wwi-02@asadatalakeinaday84.dfs.core.windows.net/top-products/*.parquet
display(df.limit(10))
```

Command executed in 2mins 35s 588ms by joel on 11-26-2020 00:53:24.571 -05:00

Job execution Succeeded **Spark** 2 executors 8 cores View in monitoring Open Spark UI

View: Table | Chart

visitorId	productId	itemsPurchased...	preferredProduc...	userId	isTopProduct	isPreferredProd...
	2717		2717	148	false	true
	4002		4002	148	false	true
	1716		1716	148	false	true
	4520		4520	148	false	true
	951		951	148	false	true
	1817		1817	148	false	true
	2634		2634	463	false	true
	2795		2795	463	false	true

129

Create notebook parameters

- Move cell up
- Move cell down
- Disable output scrolling
- Hide input
- Hide output
- Toggle parameter cell (2)

M↓ ··· (1)

```
import uuid

# Generate random GUID
runId = uuid.uuid4()

```

M↓ ···

```
%%pyspark

top5ProductsOverall.write.parquet('abfss://wwi-02@YOUR_DATALAKE_NAME.dfs.core.windows.net/top5-products/' + str(runId) + '.parquet')

```

130

© Microsoft Corporation

Integrating a notebook into a pipeline

131

Controlling notebook execution in a pipeline

132

© Microsoft Corporation

Lab: Integrate data from Notebooks with Azure Data Factory or Azure Synapse Pipelines

133

Lab overview

In the lab, the students will create a notebook to query user activity and purchases that they have been made in the past 12 months. They will then add the notebook to a pipeline using the new Notebook activity and execute this notebook after the Mapping Data Flow as part of their orchestration process. While configuring this, the students will implement parameters to add dynamic content in the control flow and validate how the parameters can be used.

Lab objectives

After completing this lab, you will be able to:

Integrate data from Notebooks with Azure Data Factory or Azure Synapse Pipelines

134

© Microsoft Corporation

135

Microsoft Azure

Module 08 : End-to-end security with Azure Synapse Analytics

136

© Microsoft Corporation

Lesson 01: Secure a data warehouse in Azure Synapse Analytics

137

Secure a data warehouse in Azure Synapse Analytics

- Network security
- Identity and access management
- Managing sensitive data
- Encryption capabilities built into Azure

138

© Microsoft Corporation

Network security

Securing your network from attacks and unauthorized access is an important part of any architecture

Internet protection	Firewalls	DDoS protection	Network security groups
Assess the resources that are internet-facing, and to only allow inbound and outbound communication where necessary. Make sure you identify all resources that are allowing inbound network traffic of any type	To provide inbound protection at the perimeter, there are several choices: • Azure Firewall • Azure Application Gateway • Azure Storage Firewall	The Azure DDoS Protection service protects your Azure applications by scrubbing traffic at the Azure network edge before it can impact your service's availability	Network Security Groups allow you to filter network traffic to and from Azure resources in an Azure virtual network. An NSG can contain multiple inbound and outbound security rules

139

Identity and access

Authentication
This is the process of establishing the identity of a person or service looking to access a resource. Azure Active Directory is a cloud-based identity service that provide this capability

Authorization
This is the process of establishing what level of access an authenticated person or service has. It specifies what data they're allowed to access and what they can do with it. Azure Active Directory also provides this capability

Azure Active Directory features

Single sign-on	Apps & device management	Identity services
Enables users to remember only one ID and one password to access multiple applications	You can manage your cloud and on-premises apps and devices and the access to your organizations resources	Manage Business to business (B2B) identity services and Business-to-Customer (B2C) identity services

140

© Microsoft Corporation

Lesson 02: Configure and manage secrets in Azure Key Vault

142

Configure and manage secrets in Azure Key Vault

Azure Key Vault protects

1. Secrets
2. Keys
3. Certificates

Data Engineers are typically concerned with accessing the data contained in Key Vault to apply to linked services

143

© Microsoft Corporation

Lesson 03: Implement compliance controls for sensitive data

144

Managing sensitive data

Column level security

```
GRANT SELECT ON wwi_security.Sale([ProductID], [Analyst], [Product], [
CampaignName],[Quantity], [Region], [State], [City], [RevenueTarget])
TO DataAnalystMiami;
```

Row level security

```
CREATE FUNCTION wwi_security.fn_securitypredicate(@Analyst AS sysname)
    RETURNS TABLE
WITH SCHEMABINDING
AS
    RETURN SELECT 1 AS fn_securitypredicate_result
    WHERE @Analyst = USER_NAME() OR USER_NAME() = 'CEO'
GO

CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE wwi_security.fn_securitypredicate(Analyst)
ON wwi_security.Sale
WITH (STATE = ON);

GRANT SELECT ON wwi_security.Sale TO CEO, DataAnalystMiami
, DataAnalystSanDiego;
```

145

© Microsoft Corporation

Implement compliance controls for sensitive data

Dynamic data masking

146

Lab: Run interactive queries using Azure Synapse Analytics serverless SQL pools

147

© Microsoft Corporation

148

Microsoft Azure

Module 09 : Support Hybrid Transactional Analytical Processing (HTAP) with Azure Synapse Link

149

© Microsoft Corporation

Lesson 01: Design hybrid transactional and analytical processing using Azure Synapse Analytics

150

Design hybrid transactional and analytical processing using Azure Synapse Analytics

151

© Microsoft Corporation

Lesson 02: Configure Azure Synapse Link with Azure Cosmos DB

152

Configure Azure Synapse Link with Azure Cosmos DB

153

© Microsoft Corporation

Lesson 03: Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics

154

Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics

- Step 1: Load the data in Spark
- Step 2: Create a base DataFrame
- Step 3: Flatten JSON data
- Step 4: Create the final DataFrame

155

© Microsoft Corporation

Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics (continued #1)

Step 1:
Load data to DataFrame

156

Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics (continued #2)

Step 2:
Create a base DataFrame

157

© Microsoft Corporation

Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics (continued #3)

Step 3:
Flattening JSON data

158

Query Azure Cosmos DB with Apache Spark for Azure Synapse Analytics (continued #4)

Step 4:
Creating the final DataFrame

159

© Microsoft Corporation

Lesson 04: Query Azure Cosmos DB with SQL Serverless for Azure Synapse Analytics

160

Query Azure Cosmos DB with SQL Serverless for Azure Synapse Analytics

Step 1:
Create a View

```
15
16  CREATE VIEW UserProfileHTAP
17  AS
18  SELECT
19      *
20  FROM OPENROWSET(
21      'CosmosDB',
22      N'account=asacosmosdbinaday84;database=CustomerProfile;key=1kxCTXbqWQ8wf8ojuzVQjCbFuAsoV6rlMR7KyD
23      UserProfileHTAP
24  )
25  WITH (
26      userId bigint,
27      cartId varchar(50),
28      preferredProducts varchar(max),
29      productReviews varchar(max)
30  ) AS profiles
31  CROSS APPLY OPENJSON (productReviews)
32  WITH (
33      productId bigint,
34      reviewText varchar(1000)
35  ) AS reviews
36  GO
```

Results | Messages

11:46:22 AM Started executing query at Line 1

(Changed database context to 'Profiles'. Changed database context to 'master'.)
Total execution time: 00:00:04.304

00:00:04 Query executed successfully.

161

© Microsoft Corporation

Lab: Support Hybrid Transactional Analytical Processing (HTAP) with Azure Synapse Link

162

Lab overview

This lab teaches you how Azure Synapse Link enables seamless connectivity of an Azure Cosmos DB account to an Azure Synapse workspace. You will understand how to enable and configure Synapse link, then how to query the Azure Cosmos DB analytical store using Apache Spark and SQL Serverless.

Lab objectives

After completing this lab, you will be able to:

- Configure Azure Synapse Link with Azure Cosmos DB
- Query Azure Cosmos DB with Apache Spark for Synapse Analytics
- Query Azure Cosmos DB with serverless SQL pool for Azure Synapse Analytics

163

© Microsoft Corporation

164

Microsoft Azure

Module 10 : Real-time Stream Processing with Stream Analytics

165

© Microsoft Corporation

Lesson 01: Enable reliable messaging for Big Data applications using Azure Event Hubs

166

Azure Event Hubs

Azure Event Hubs is a highly scalable publish-subscribe service that can ingest millions of events per second and stream them into multiple applications

167

© Microsoft Corporation

Create an Event Hub

Create an event hub namespace

1. In the Azure portal, select **+ Create a Resource,** type **Event Hubs,** and then select **Event Hubs** from the resulting search. Then select **Create**
2. Provide a name for the event hub, and then create a resource group. Specify **xx-name-eh** and **xx-name-rg** respectively, XX- represent your initials to ensure uniqueness of the Event Hub name and Resource Group name
3. Click the checkbox to **Pin to the dashboard**, then select the **Create** button

Create an event hub

1. After the deployment is complete, click the **xx-name-eh** event hub on the dashboard
2. Then, under **Entities**, select **Event Hubs**
3. To create the event hub, select the **+ Event Hub** button. Provide the name **socialstudy-eh**, and then select **Create**
4. To grant access to the event hub, we need to create a shared access policy. Select the **socialstudy-eh** event hub when it appears, and then, under **Settings**, select **Shared access policies**
5. Under **Shared access policies**, create a policy with **MANAGE** permissions by selecting **+ Add**. Give the policy the name of **xx-name-eh-sap**, check **MANAGE**, and then select **Create**
6. Select your new policy after it has been created, and then select the copy button for the **CONNECTION STRING – PRIMARY KEY** entity
7. Paste the **CONNECTION STRING – PRIMARY KEY** entity into Notepad, this is needed later in the exercise
8. Leave all windows open

168

Configure applications to use Event Hubs

169

© Microsoft Corporation

Evaluating the performance of Event Hubs

170

Lesson 02: Work with data streams by using Azure Stream Analytics

171

© Microsoft Corporation

What are data streams

Data streams:
In the context of analytics, data streams are event data generated by sensors or other sources that can be analyzed by another technology

Data stream processing approach:
There are two approaches. Reference data is streaming data that can be collected over time and persisted in storage as static data. In contrast, streaming data have relatively low storage requirements. And run computations in sliding windows

Data streams are used to:

Analyze data:	Understand systems:	Trigger actions:
Continuously analyze data to detect issues and understand or respond to them	Understand component or system behavior under various conditions to fuel further enhancements of said system	Trigger specific actions when certain thresholds are identified

172

Event processing

The process of consuming data streams, analyzing them, and deriving actionable insights out of them is called Event Processing and has three distinct components:

Event producer	Examples include sensors or processes that generate data continuously such as a heart rate monitor or a highway toll lane sensor
Event processor	An engine to consume event data streams and deriving insights from them. Depending on the problem space, event processors either process one incoming event at a time (such as a heart rate monitor) or process multiple events at a time (such as a highway toll lane sensor)
Event consumer	An application which consumes the data and takes specific action based on the insights. Examples of event consumers include alert generation, dashboards, or even sending data to another event processing engine

173

© Microsoft Corporation

Processing events with Azure Stream Analytics

Microsoft Azure Stream Analytics is an event processing engine. It enables the consumption and analysis of high volumes of streaming data in real time

Source	Ingestion	Analytical engine	Destination
Sensors Systems Applications	Event Hubs IoT Hubs Azure Blob Store	Stream Analytics Query Language .NET SDK	Azure Data Lake Cosmos DB SQL Database Blob Store Power BI

174

Lesson 03: Transform data by using Azure Stream Analytics

175

© Microsoft Corporation

Create Stream Analytics service

Job name	Subscription	Resource group	Location

176

Create a Stream Analytics Job input

177

© Microsoft Corporation

Create a Stream Analytics Job output

178

Transform data by using Azure Stream Analytics

- Declarative SQL like language
- Contains testing capabilities
- Performs aggregations over windows of time

179

© Microsoft Corporation

Using Windowing functions with Azure Stream Analytics

Performing operations on the data contained in temporal windows is a common pattern. Stream Analytics has native support for windowing function including

Tumbling	Hopping	Sliding	Session
Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them. For example, tell me the count of tweets per time zone every 10 seconds	Hopping window functions hop forward in time by a fixed period. For example, every 5 seconds, give me the count of tweets for the last 10 seconds	Sliding windows, output events only for points in time when the content of the window actually changes. For example, give a count of all tweets which are tweeted more than 10 times on a given topic	Session window functions group events that arrive at similar times, filtering out periods of time where there is no data. For example, give the count of tweets that occur within 5 minutes of each other

180

Windowing functions examples with Azure Stream Analytics

Tumbling	Hopping	Sliding	Session
SELECT Timezone, Count(*) as Count FROM TwitterStream TIMESTAMP BY CreatedAt GROUP BY Timezone TUMBLINGWINDOW (second, 10)	SELECT Timezone, Count(*) as TotalTweet FROM TwitterStream TIMESTAMP BY CreatedAt GROUP BY Topic HOPPINGWINDOW (second, 10, 5)	SELECT Topic, Count(*) FROM TwitterStream TIMESTAMP BY CreatedAt GROUP BY Topic SLIDINGWINDOW (second, 10) HAVING COUNT(*) > 10	SELECT Timezone, Count(*) FROM TwitterStream TIMESTAMP BY CreatedAt GROUP BY Topic SESSIONWINDOW (minute 5, 10)

181

© Microsoft Corporation

Start a Stream Analytics Job

182

Lab: Real-time Stream Processing with Stream Analytics

183

© Microsoft Corporation

Lab overview

This lab teaches you how to process streaming data with Azure Stream Analytics. You will ingest vehicle telemetry data into Event Hubs, then process that data in real time, using various windowing functions in Azure Stream Analytics. You will output the data to Azure Synapse Analytics. Finally, you will learn how to scale the Stream Analytics job to increase throughput.

Lab objectives

After completing this lab, you will be able to:

- Use Stream Analytics to process real-time data from Event Hubs
- Use Stream Analytics windowing functions to build aggregates and output to Synapse Analytics
- Scale the Azure Stream Analytics job to increase throughput through partitioning
- Repartition the stream input to optimize parallelization

184

185

© Microsoft Corporation

Microsoft Azure

Module 11: Create a Stream Processing Solution with Event Hubs and Azure Databricks

186

Lesson 01: Understand the key features and uses of Structured Streaming

187

© Microsoft Corporation

Understand the key features and uses of Structured Streaming

Fast:
Processes millions of data in a matter of seconds.

Scalable:
Enables you to auto-scale

Fault-tolerant:
Structured Streaming automatically checkpoints the state data to fault-tolerant storage

Integration with other Cloud Services:
Can integrate with a variety of Azure data platform services such as Azure Event Hubs

188

Lesson 02: Stream data from a file and write it out to a distributed file system and connect to Event Hubs to read and write streams

189

© Microsoft Corporation

Stream data from a file and write it out to a distributed file system and connect to Event Hubs to read and write streams

190

Configuring a file stream

Cmd 6

```
# Here we define the schema using a DDL-formatted string (the SQL Data Definition Language).
dataSchema = "Recorded_At timestamp, Device string, Index long, Model string, User string, _corrupt_record String, gt string, x double, y double, z double"

dataPath = "dbfs:/mnt/training/definitive-guide/data/activity-data-stream.json"
initialDF = (spark
  .readStream                           # Returns DataStreamReader
  .option("maxFilesPerTrigger", 1)      # Force processing of only 1 file per trigger
  .schema(dataSchema)                   # Required for all streaming DataFrames
  .json(dataPath)                       # The stream's source directory and file type
)
```

Cmd 10

```
streamingDF = (initialDF
  .withColumnRenamed("Index", "User_ID")  # Pick a "better" column name
  .drop("_corrupt_record")                # Remove an unnecessary column
)
```

191

© Microsoft Corporation

Lesson 03: Use sliding windows to aggregate over chunks of data rather than all data

192

Use sliding windows to aggregate over chunks of data rather than all data

193

© Microsoft Corporation

Lesson 04: Apply watermarking

194

Apply watermarking

Prevent Data Build Up
Over time, aggregated data will build up in the driver

Prevents Long time running Jobs
Building up an unbounded set of windows, causing hit of resource limits

195

© Microsoft Corporation

Lab: Create a Stream Processing Solution with Event Hubs and Azure Databricks

196

Lab overview

This lab teaches you how to ingest and process streaming data at scale with Event Hubs and Spark Structured Streaming in Azure Databricks. You will learn the key features and uses of Structured Streaming. You will implement sliding windows to aggregate over chunks of data and apply watermarking to remove stale data. Finally, you will connect to Event Hubs to read and write streams.

Lab objectives

After completing this lab, you will be able to:

- Understand the key features and uses of Structured Streaming
- Stream data from a file and write it out to a distributed file system and connect to Event Hubs to read and write streams
- Use sliding windows to aggregate over chunks of data rather than all data
- Apply watermarking to remove stale data

197

© Microsoft Corporation

198

Microsoft Azure

… The End …

© Copyright Microsoft Corporation. All rights reserved.

199

© Microsoft Corporation