Microsoft

DP-200T01: Azure for the Data Engineer

Agenda

- L01 – Explain the evolving world of data
- L02 - Survey the services in the Azure Data Platform
- L03 - Identify the tasks that are performed by a Data Engineer
- L04 - Describe the use cases for the cloud in a case study

Lesson 01

The Evolving World of Data

Lesson Objectives

- Data abundance
- Differences between on-premises and cloud data technologies
- How the role of the data professional is changing in organizations
- Identify use cases impacted by these changes

Data abundance

Processes

Businesses are tasked to store, interpret, manage, transform, process, aggregate and report on data

Consumers

There are a wider range of consumers using different types of devices to consume or generate data

Variety

There's a wider variety of data types that need to be processed and stored

Responsibilities

A data engineers role is responsible for more data types and technologies

Technologies

Microsoft Azure provides a wide set of tools and technologies

On-premises versus cloud technologies

Data engineering job responsibilities

Use cases for the cloud

Here are some examples of industries making use of the cloud

Web retail	Using Azure Cosmos DB's multi-master replication model along with Microsoft's performance commitments, Data Engineers can implement a data architecture to support web and mobile applications that achieve less than a 10-ms response time anywhere in the world
Healthcare	Azure Databricks can be used to accelerate big data analytics and artificial intelligence (AI) solutions. Within the healthcare industry, it can be used to perform genome studies or pharmacy sales forecasting at petabyte scale
IoT scenarios	Hundreds of thousands of devices have been designed and sold to generate sensor data known as Internet of Things (IoT) devices. Using technologies like Azure IoT Hub, Data Engineers can easily design a data solution architecture that captures real-time data

Review Questions

- Q01 – What data processing framework will be mainly used by Data Engineers to ingest data into cloud data platforms on Azure?
- A01 – **Extract, Load and Transform (ELT)**

- Q02 - What type of data can have its own schema at defined at query time?
- A02 – **Unstructured data**

- Q03 - Duplicating customers content for redundancy and meeting service level agreements (SLA) in Azure meets which cloud technical requirement?
- A03 – **High availability**

Lesson 02

Survey the Services in the Azure Data Platform

Lesson Objectives

- The differences between structured and unstructured data
- Azure Storage
- Azure Data Lake Storage
- Azure Databricks
- Azure Cosmos DB
- Azure SQL Database
- Azure SQL Data Warehouse
- Azure Stream Analytics
- Additional Azure Data Platform Services

Structured versus unstructured data

There are three broad types of data and Microsoft Azure provides many data platform technologies to meet the needs of the wide varieties of data

Structured	Semi- Structured	Unstructured
Structured data is data that adheres to a schema, so all of the data has the same fields or properties. Structured data can be stored in a database table with rows and columns	Semi-structured data doesn't fit neatly into tables, rows, and columns. Instead, semi-structured data uses _tags_ or _keys_ that organize and provide a hierarchy for the data	Unstructured data encompasses data that has no designated structure to it. Known as No-SQL., there are four types of No-SQL databases: • Key Value Store • Document Database • Graph Databases • Column Base

What to use for Data

Storage Account

Module 02

- 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

Module 02

- 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

Module 03

- **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 CosmosDB

Module 04

- 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

Module 05

- 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**

What to use for Data

Azure Synapse Analytics

Module 05

- 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

Module 06

- 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

Module 07

- 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 Data Catalog

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

Review Questions

- Q01 - What data platform technology is a globally distributed, multi-model database that can offer sub second query performance?
- A01 – Azure Cosmos DB

- Q02 - Which of the following is the cheapest data store to use when you want to store data without the need to query it?
- A02 – Azure Storage Account

- Q03 - Which Azure Service would be used to store documentation about a data source?
- A03 – Azure Data Catalog

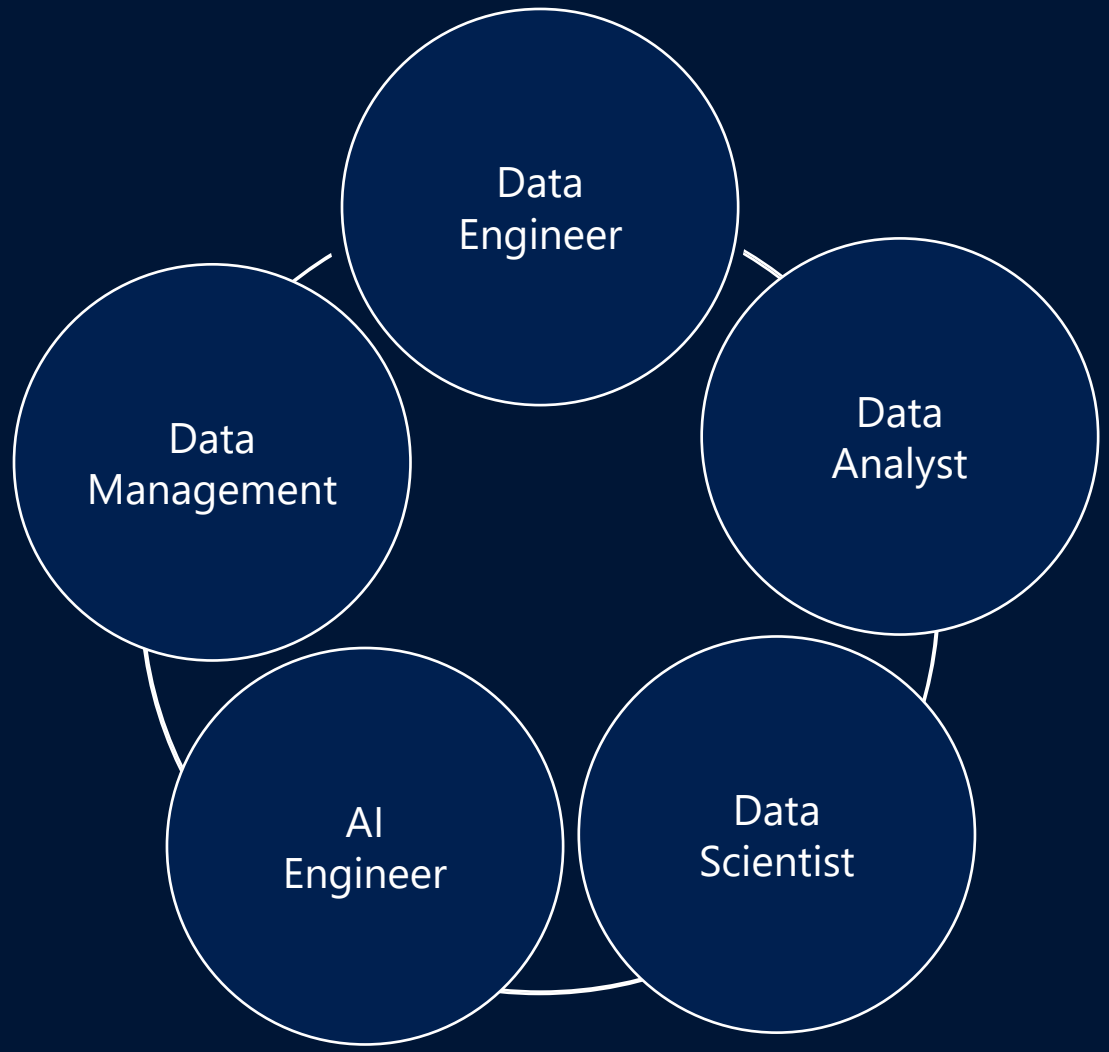
Lesson 03

Identify the Tasks Performed by a Data Engineer

Lesson Objectives

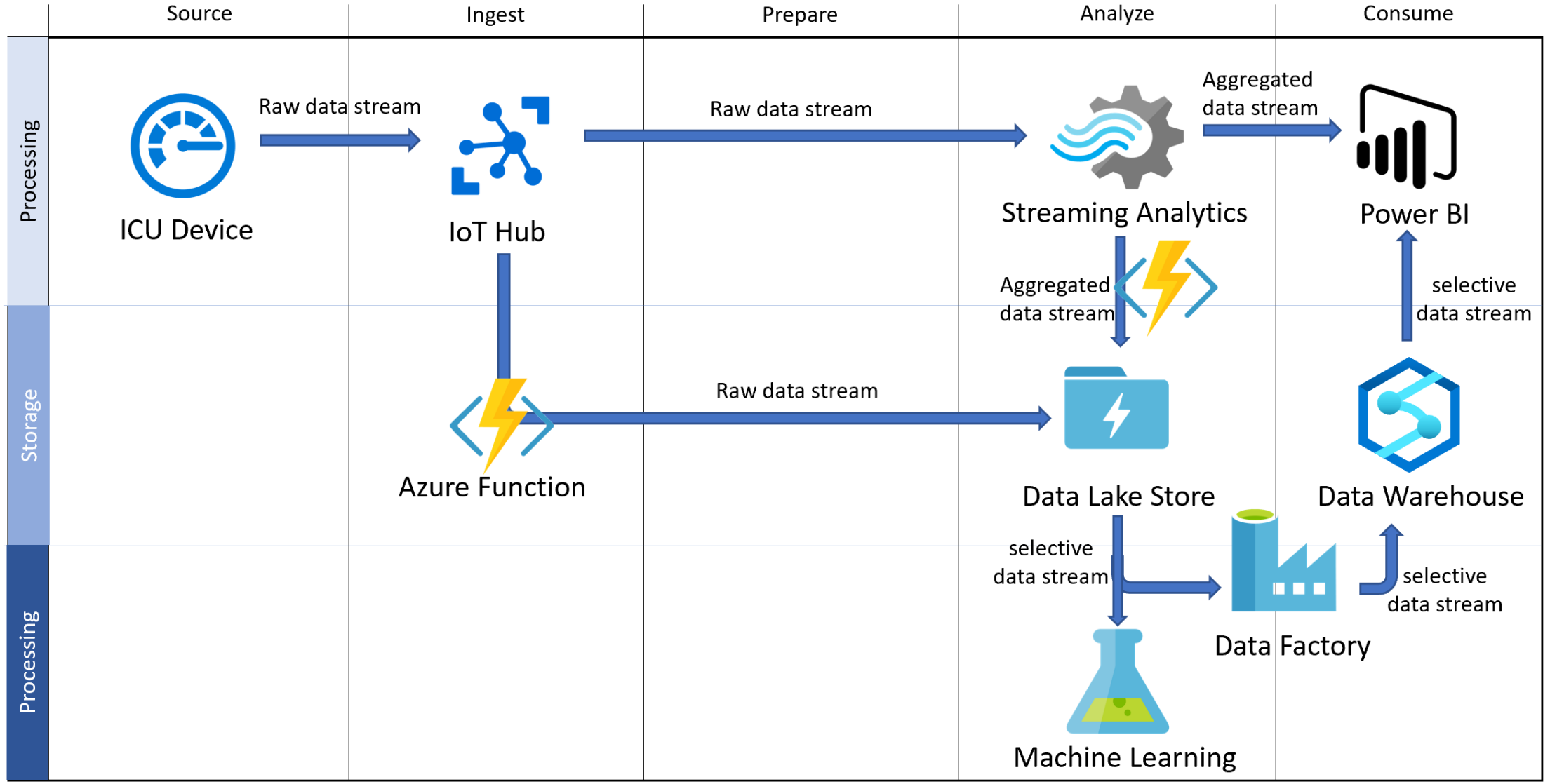
- List the new roles of modern data projects
- Outline data engineering practices
- Explore the high-level process for architecting a data engineering project

Roles in Data Projects



Data Engineering Practices

Architecting Projects – an example



Review Questions

- Q01 – Which role works with services such as Cognitive Services, Cognitive Search, and the Bot Framework?
- A01 – AI Engineer

- Q02 - Which Azure Data Platform technology is commonly used to process data in an ELT framework?
- A02 – Azure Data Factory

Lesson 04

Course Case Study

Lesson Objectives

- Read the course case study

Course case study: AdventureWorks Cycles

Read the case study

In this section of the course, the instructor will either:

- Allocate you 10 minutes to read through the case study.

  Or

- Spend 10 minutes walking through the case study with you as a group

Note:
This case study will be used in labs across the entire course. Each lab will drill down more into the detail of what is required as you perform each lab.

Review Questions

- Q01 – Which requirement is likely to be the easiest to implement from a data engineering perspective?
- A01 – AdventureWorks Website

- Q02 - Which data platform technology could be used to implement the predictive analytics capabilities that AdventureWorks desires?
- A02 – Azure Databricks

- Q03 - Which data platform technology could be used to help AdventureWorks scale globally?
- A03 – Azure Cosmos DB

Lab: Azure for the Data Engineer

Lab overview

The students will take the information gained in the lessons and from the case study to scope out the deliverables for a digital transformation project within AdventureWorks. They will first identify how the evolving use of data has presented new opportunities for the organization. The students will also explore which Azure Data Platform services can be used to address the business needs and define the tasks that will be performed by the data engineer. Finally, students will finalize the data engineering deliverables for AdventureWorks.

Lab objectives

After completing this lab, you will be able to:

1. Identify the evolving world of data within AdventureWorks.
2. Determine the Azure Data Platform services to use for AdventureWorks.
3. Identify the tasks to be performed by the Data Engineer.
4. Finalize the data engineering deliverables for AdventureWorks.

Lab scenario

You have been hired as a Senior Data Engineer to advise the management and employees of AdventureWorks on a digital transformation project. AdventureWorks has been selling bicycles and bicycle parts directly to end-consumer and distributors for over a decade. In the last few years, they have observed how different industries have been taking advantage of recent developments in cloud technology to offer customers a more personalized and engaging service. They want to lead the way in their industry.

You have been asked to work with the IT department to perform a discovery workshop with the organization to identify which data platform technologies can be used to the benefit of both AdventureWorks and their customers. They are specifically wanting to make sure that the technology brings the business closer to their customers on a variety of levels. This can include making personalized offers when making purchases or using customer services, to offering telemetry information on the bikes that they use. There is also a requirement to manage an existing reporting system that is held in a data warehouse.

The first step is to assess the current state of the cloud technologies and make sure that they can perform the project by performing the following analysis:

1. Identify the evolving world of data within AdventureWorks
2. Determine the Azure Data Platform services to use for AdventureWorks
3. Identify the tasks to be performed by the Data Engineer
4. Finalize the data engineering deliverables for AdventureWorks

Lab review

- Exercise 1 – Have any data requirements been missed?
- Exercise 2 – Was there any debate in the mapping of a technology against a given data requirement?
- Exercise 3 – Are there any additional tasks that you think a Data Engineer should perform?
- Exercise 4 – Which particular technologies do you want to focus on in this course?

Module Summary

In this module, you have learned about:
- The evolving world of data.
- The services in the Azure Data Platform.
- The tasks that are performed by a Data Engineer.
- A fictitious Case Study for use in labs.

Next steps

After the course, consider visiting [the Microsoft Customer Case Study site. Use the search bar to search by an industry such as healthcare or retail, or by a technology such as Azure Cosmos DB or Stream Analytics. Read through some of the customers stories.