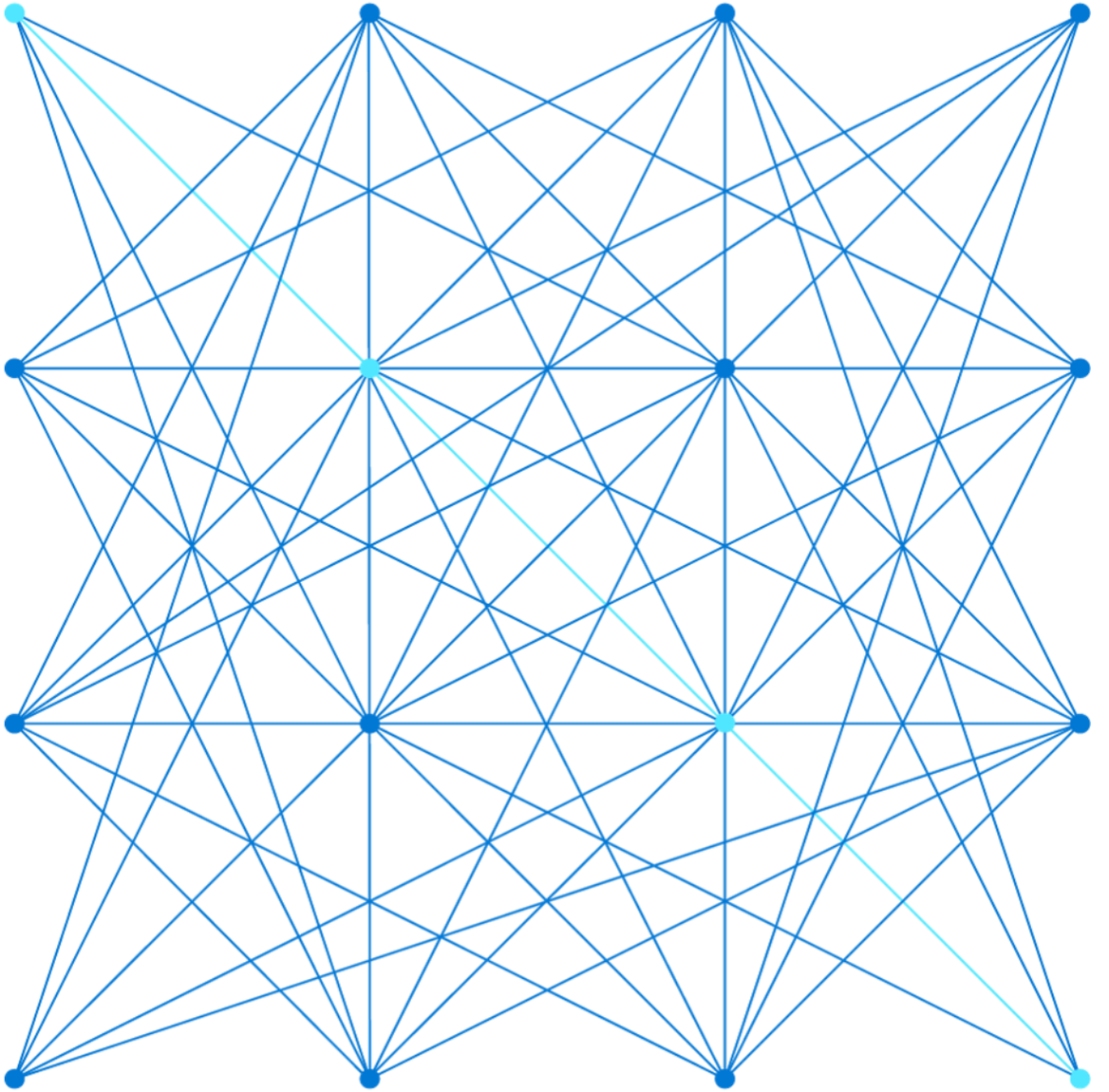
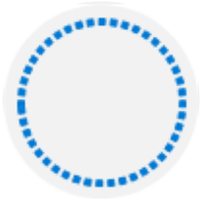
Microsoft Azure

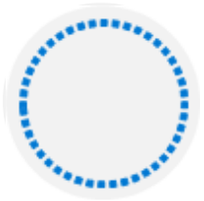
DP-203T00: Ingest and load data into the data warehouse



Agenda



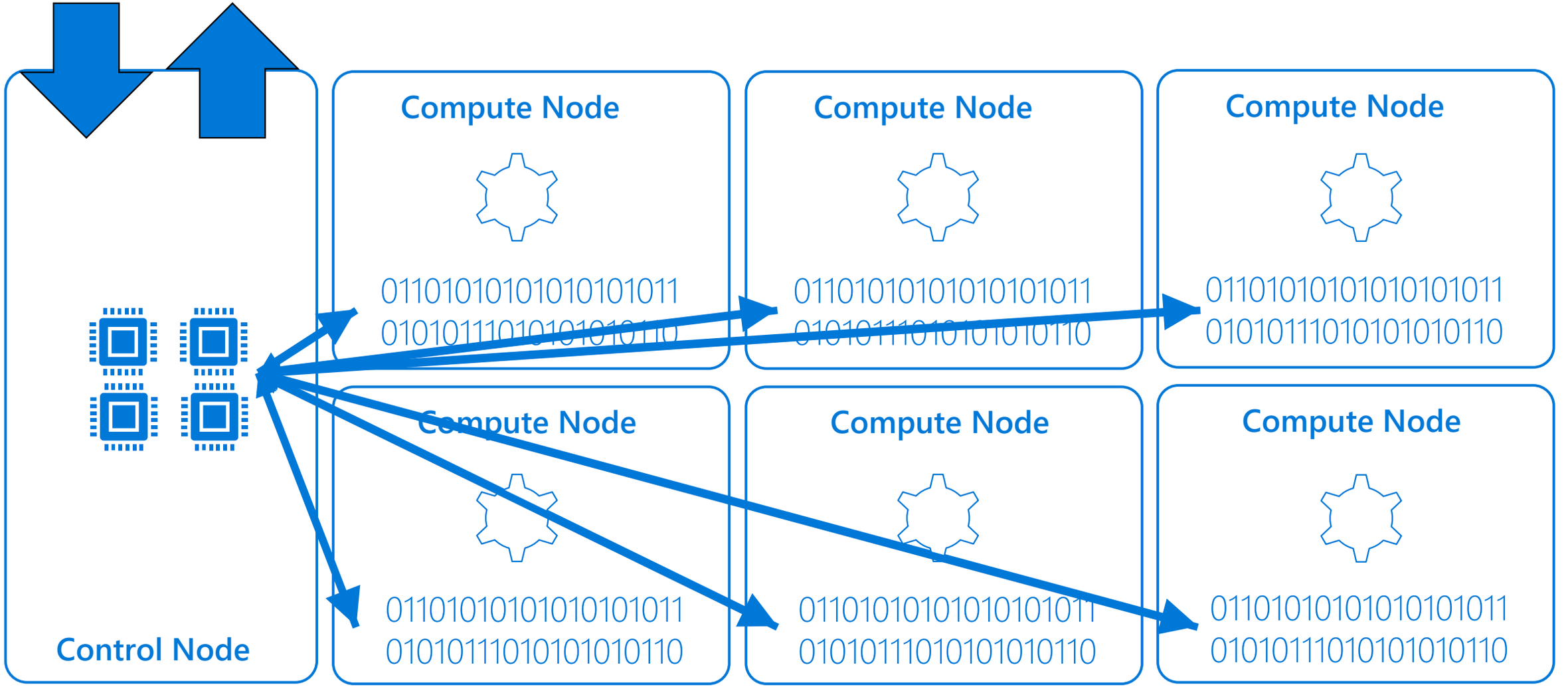
Lesson 01: Use data loading best practices in Azure Synapse Analytics



Lesson 02: Petabyte-scale ingestion with Azure Data Factory

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

Dedicated SQL Pool architecture revision



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?

Manage singleton updates

Control Node

Compute Node

01101010101010101011
01010111010101010110

Compute Node

01101010101010101011
01010111010101010110

Compute Node

01101010101010101011
01010111010101010110

Compute Node

01101010101010101011
01010111010101010110

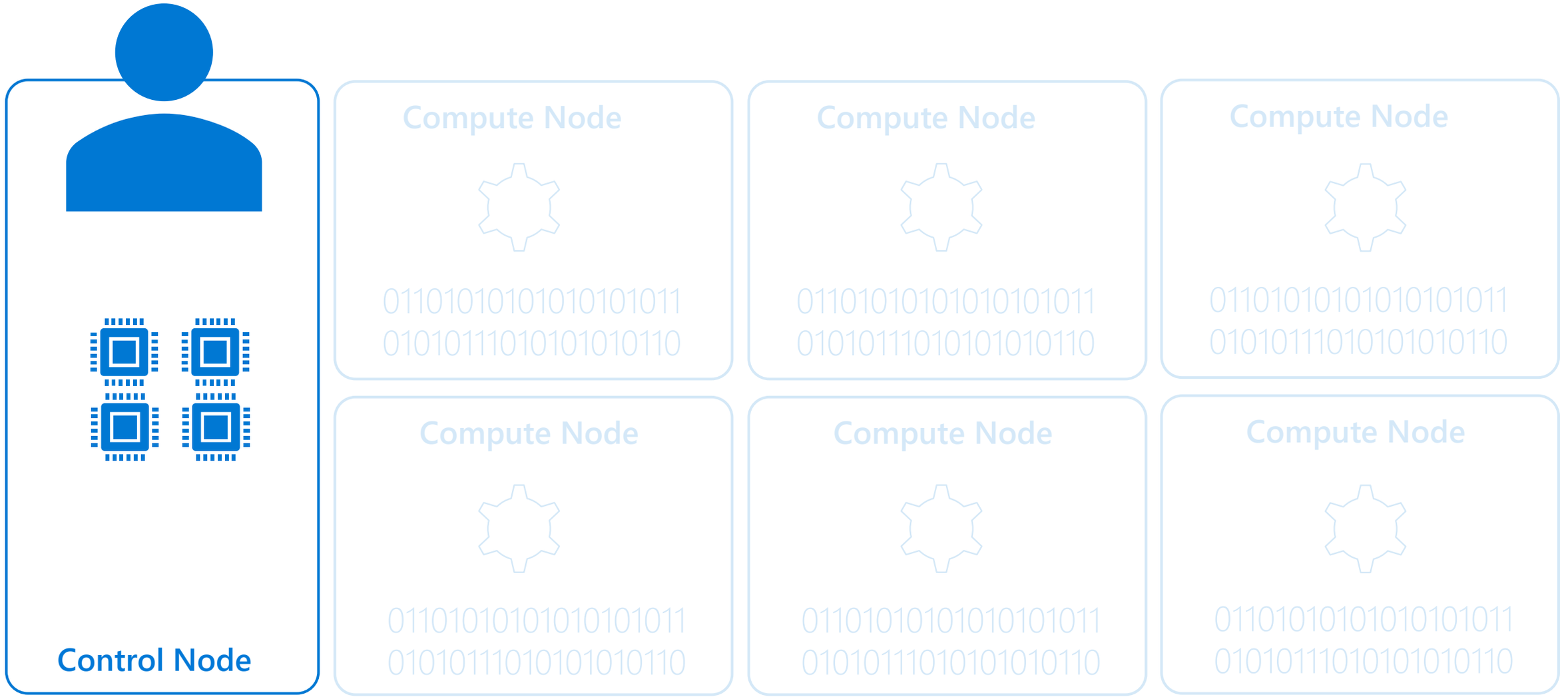
Compute Node

01101010101010101011
01010111010101010110

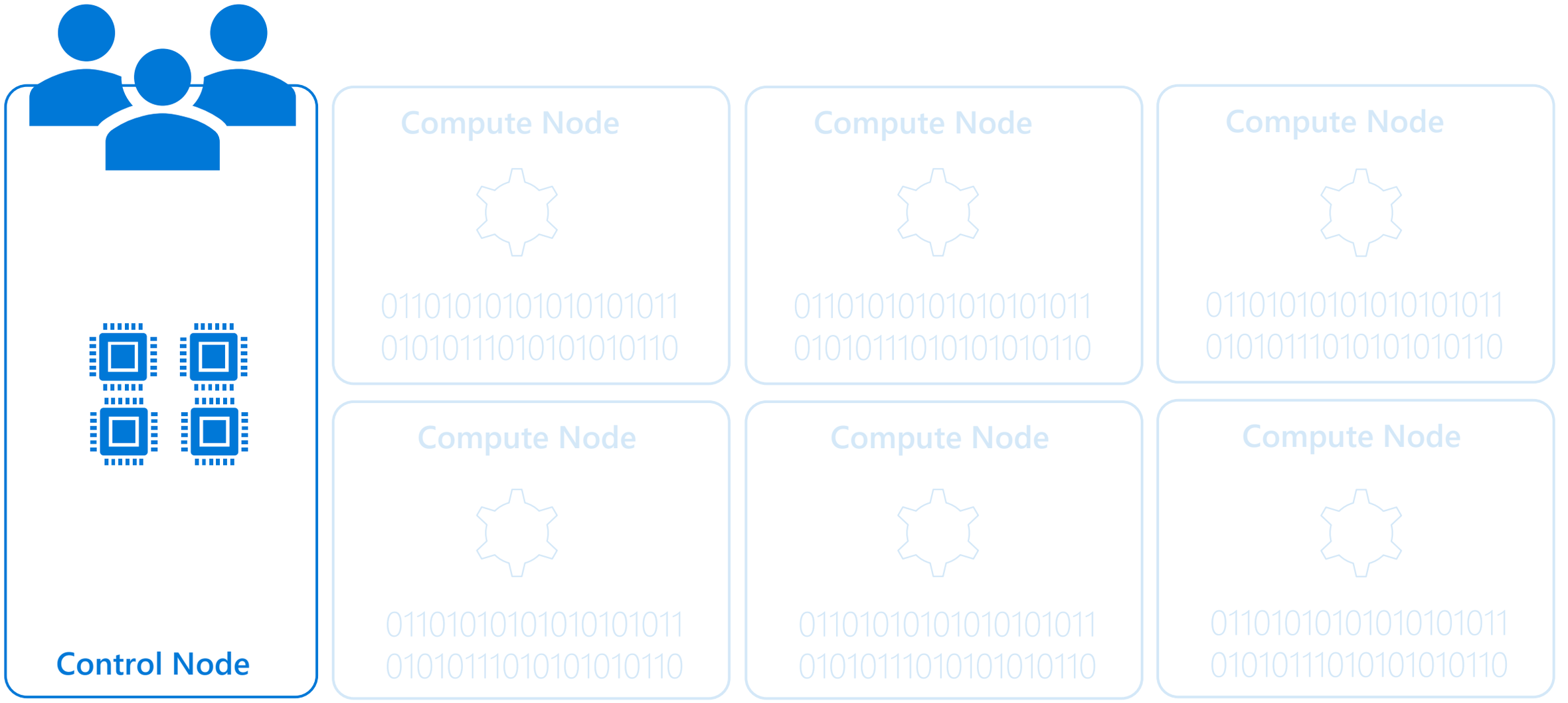
Compute Node

01101010101010101011
01010111010101010110

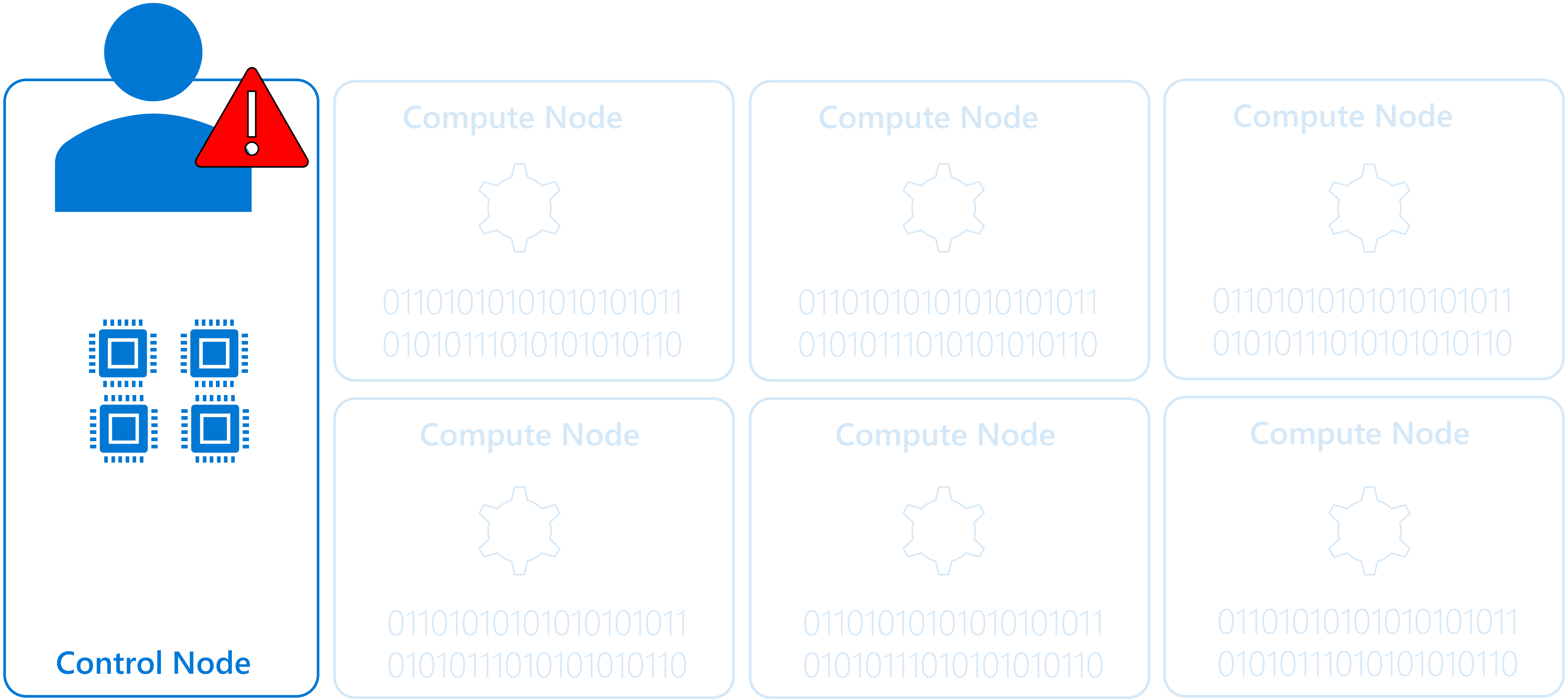
Set-up dedicated data loading accounts



Manage concurrent access to Azure Synapse Analytics

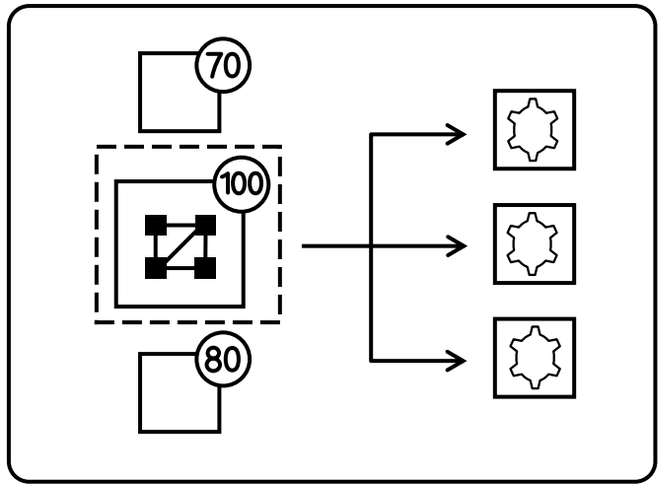


Implement Workload Management

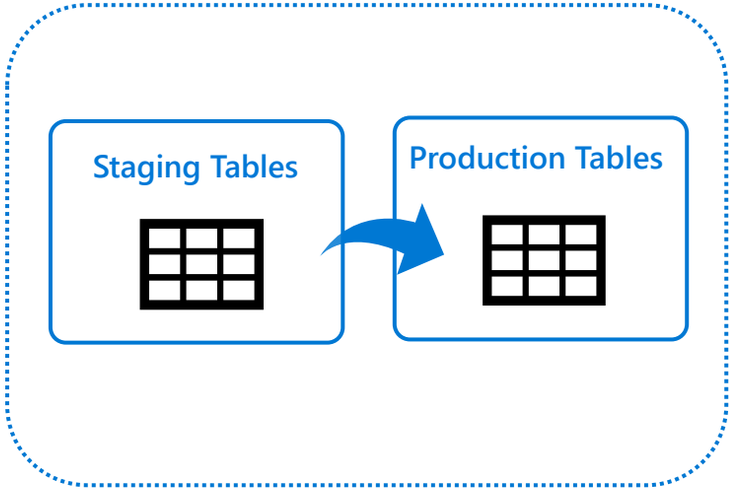


Use PolyBase, the Copy command or the Copy Activity

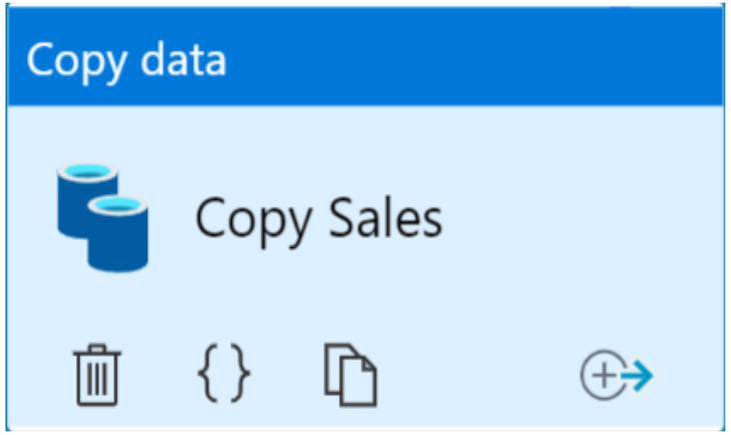
PolyBase



Copy command

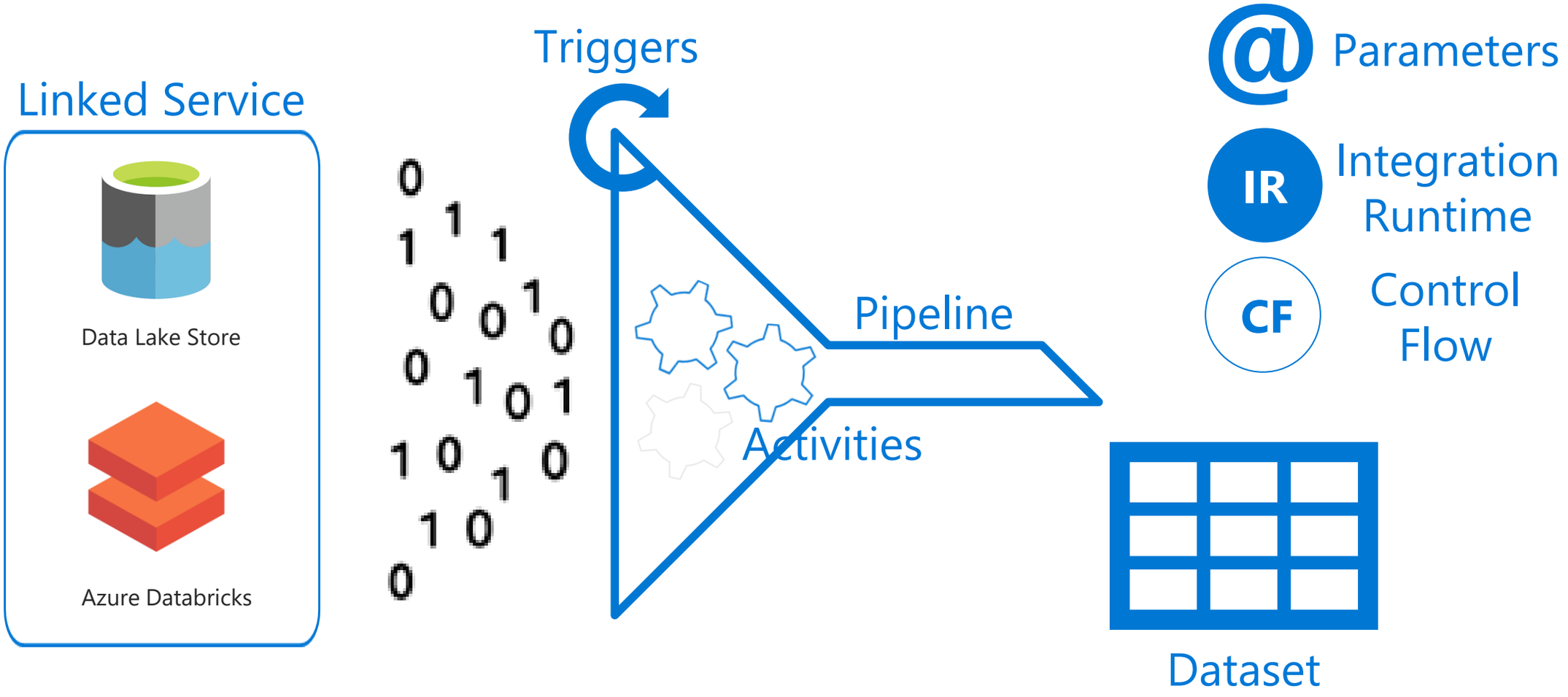


Copy data activity

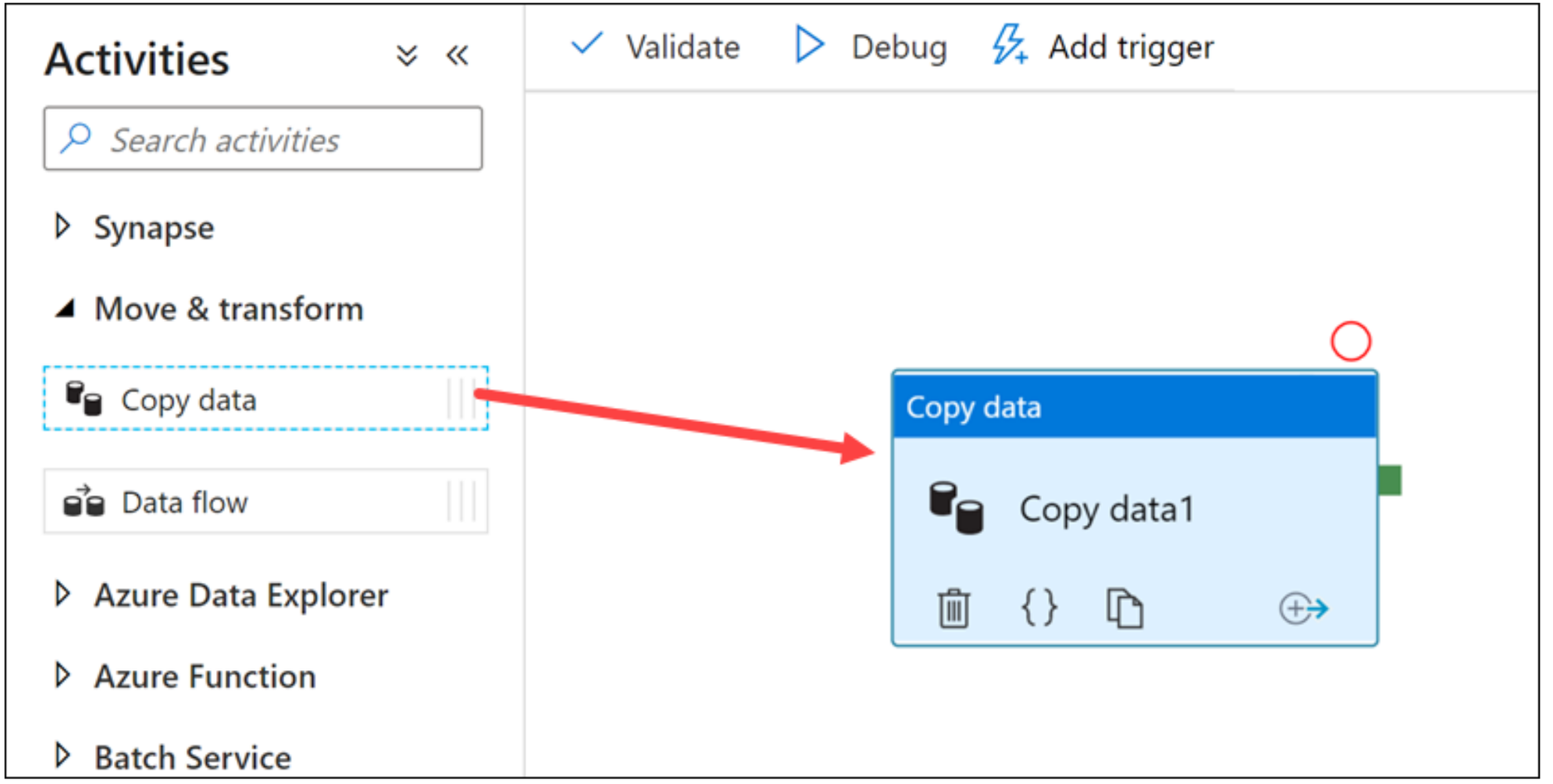


Lesson 02: Petabyte-scale ingestion with Azure Data Factory

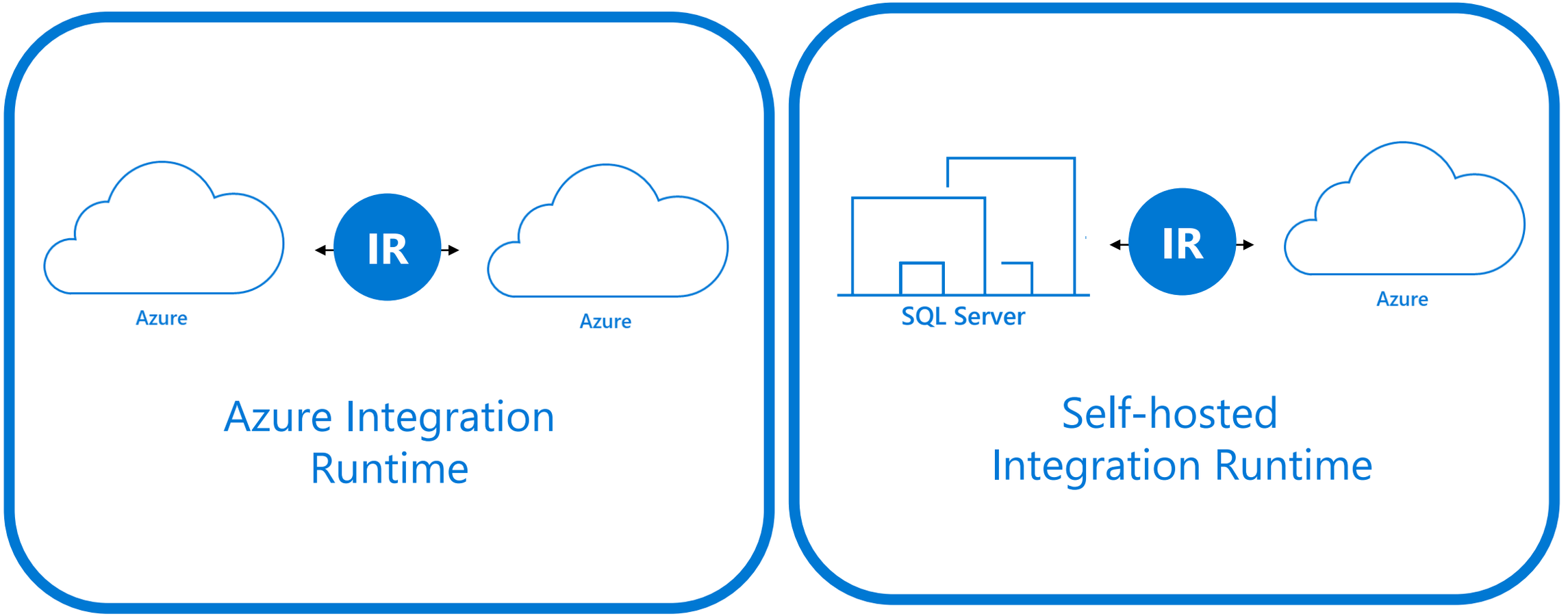
Azure Data Factory/Synapse pipeline revision



Petabyte-scale ingestion with Azure Data Factory



Understanding integration



Review questions

Q01 – Which data loading feature limits the number of resources a group of requests can consume in Azure Synapse Analytics?

A01 – Workload management

Q02 – Why should you split up one large files into smaller files when loading data into a dedicated SQL pool in Azure Synapse Analytics?

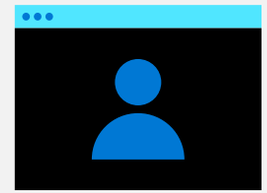
A02 – To take advantage of the Massively Parallel Processing (MPP) architecture

---

Q03 – In which section of the Data Factory designer canvass would you find the Copy Activity?

A03 – Move and Transform

Lab: Ingest and load data into the data warehouse



http://fmdk.io/dp2035

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

Lab review

Question 1 – What is the appropriate distribution to use in a staging table?

Question 2 – When creating an External Data Source, what should the Type property be set to access data in a data lake?

Question 3 – When setting file format properties for a data source in the Copy Activity, you can import the schema from a connection/store and what other option?

Question 4 – You want to truncate a staging table before loading it with data in a Copy activity. Which sink property would you set to do this?

Module summary

In this module, you have learned about:

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

Next steps

After the course, consider reading [Best practices for loading data using dedicated SQL pools in Azure Synapse Analytics] for more guidance on data ingestion and loading

Microsoft Azure