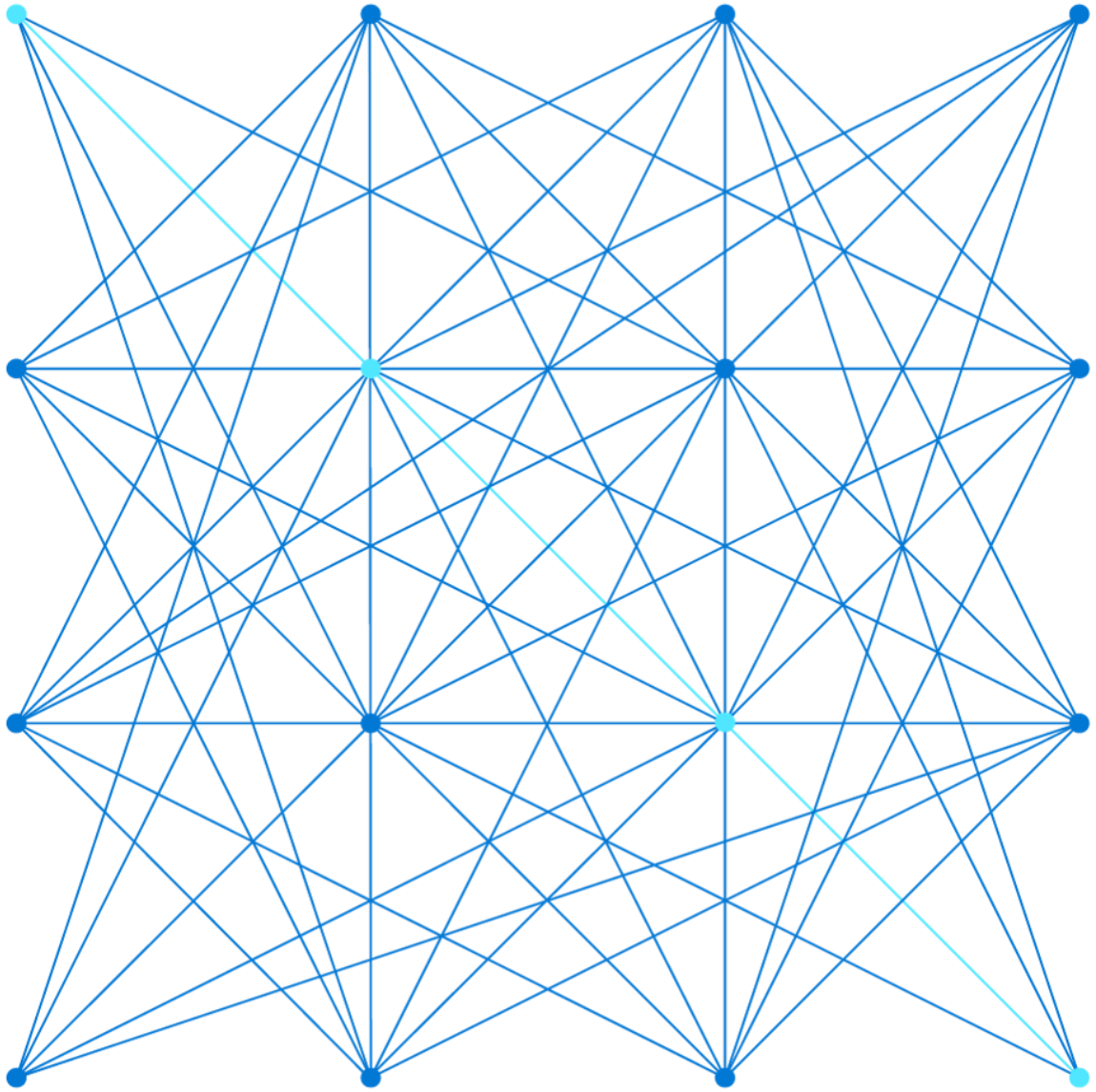
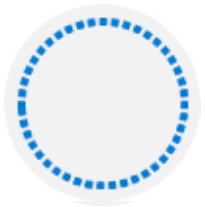
Microsoft Azure

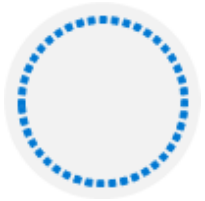
DP-203T00: Create a Stream Processing Solution with Event Hubs and Azure Databricks



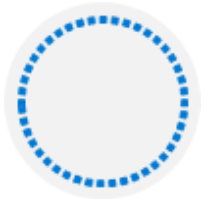
Agenda



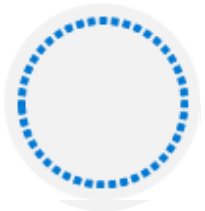
Lesson 01 – Understand the key features and uses of Structured Streaming



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



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



Lesson 04 – Apply watermarking

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

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:

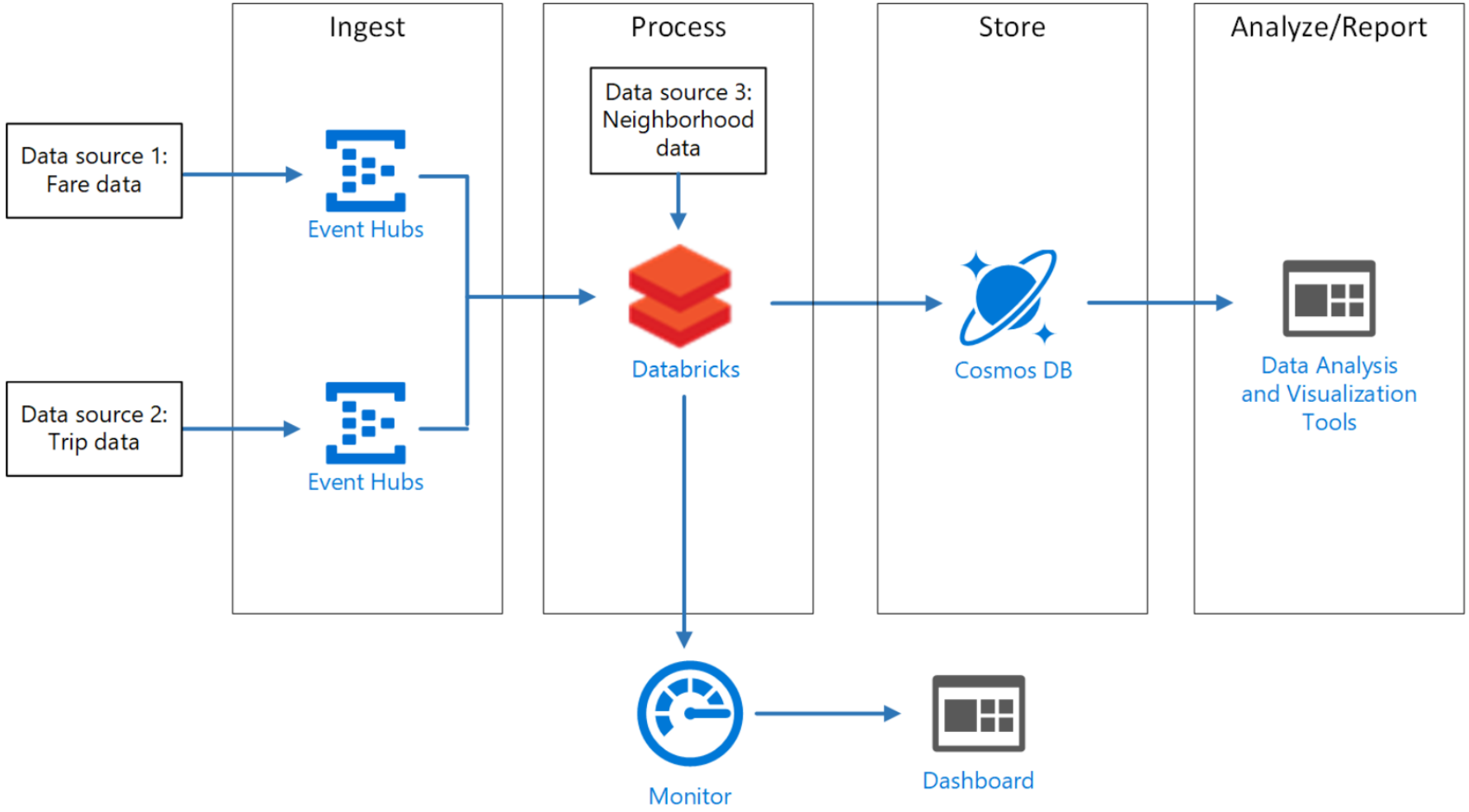
Structured Streaming automatically checkpoints the state data to fault-tolerant storage

Integration with other Cloud Services:

Can integrate with a variety of Azure data platform services such as Azure Event Hubs

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

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



Configuring a file stream

Cmd 6

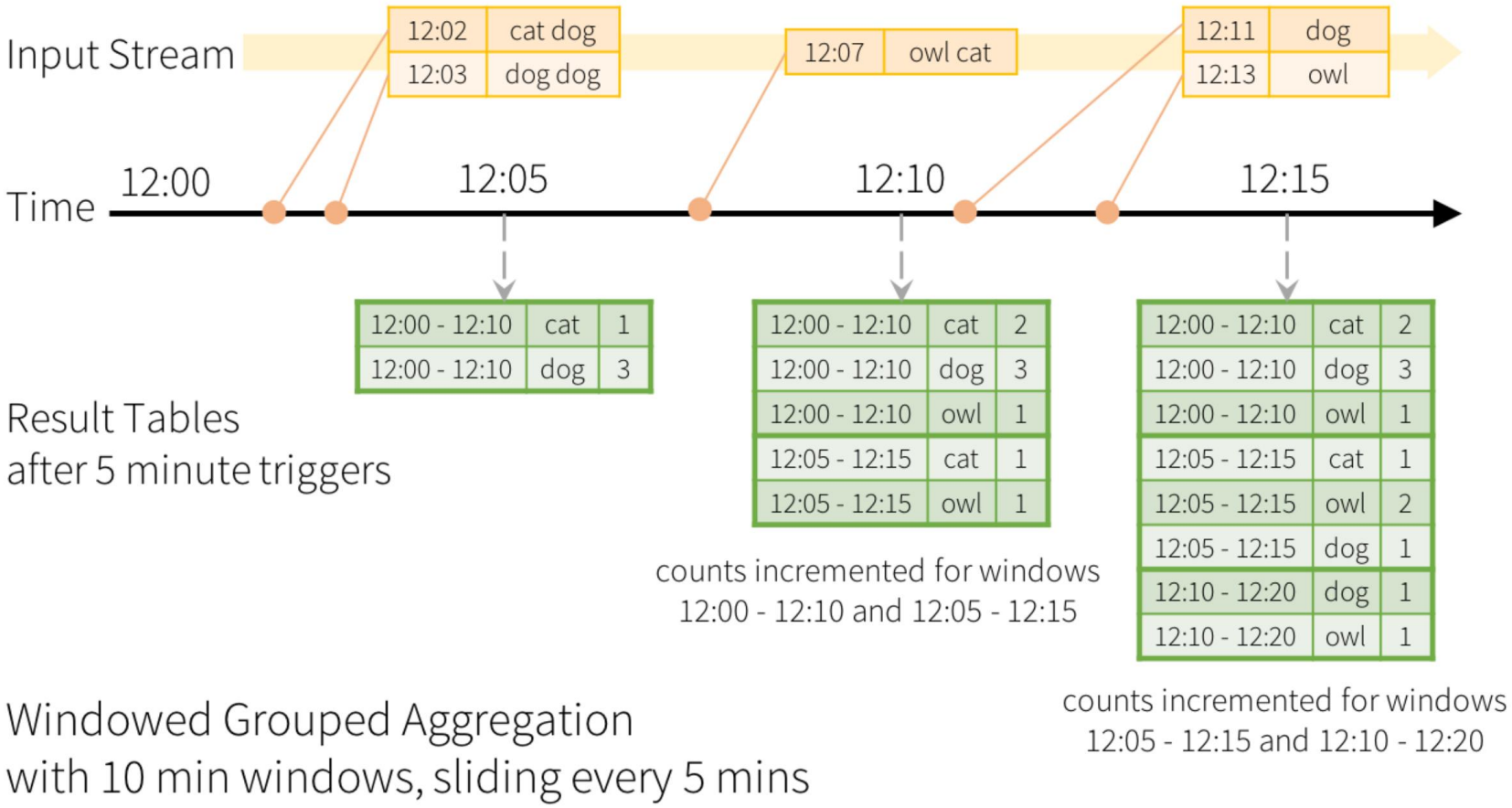
```
# Here we define the schema using a DDL-formatted string (the SQL Data Definition Language).
dataSchema = "Recorded_At timestamp, Device string, Index long, Model string, User string, _corrupt_record String, gt string, x double, y double, z double"

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
)
```

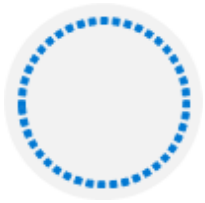
Lesson 03: Use sliding windows to aggregate over chunks of data rather than all data

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



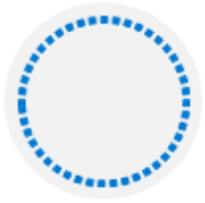
Lesson 04: Apply watermarking

Apply watermarking



Prevent Data Build Up

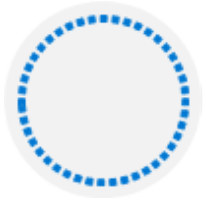
Over time, aggregated data will build up in the driver



Prevents Long time running Jobs

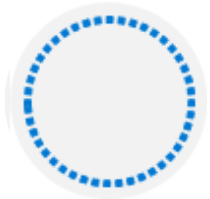
Building up an unbounded set of windows, causing hit of resource limits

Review questions



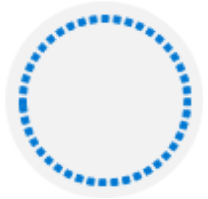
Q01 – When doing a write stream command, what does the outputMode("append") option do?

A01 – The append outputMode allows records to be added to the output sink



Q02 – In Spark Structured Streaming, what method should be used to read streaming data into a DataFrame?

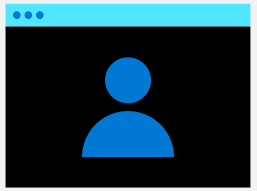
A02 – spark.readStream



Q03 – What happens if the command option("checkpointLocation", pointer-to-checkpoint directory) is not specified

A03 – When the streaming job stops, all state around the streaming job is lost, and upon restart, the job must start from scratch

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



http://fmdk.io/dp20311

Lab overview

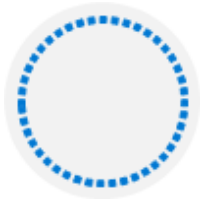
This lab teaches you how to ingest and process streaming data at scale with Event Hubs and Spark Structured Streaming in Azure Databricks. You will learn the key features and uses of Structured Streaming. You will implement sliding windows to aggregate over chunks of data and apply watermarking to remove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

Lab objectives

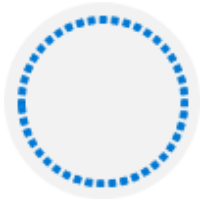
After completing this lab, you will be able to:

- 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

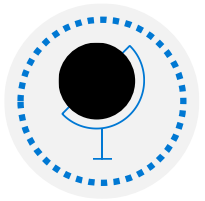
Lab review



Q01 – What is watermarking?



Q02 – What does this method do SparkSession.readStream?



Q03 – What is windowing?

Module summary

In this module, you have learned about:

- Azure Event Hubs
- Azure Databricks
- Sliding Windows
- Watermarking
- Structured Streaming

Next steps

After the course, consider visiting the website that explores [structured streaming] patterns with Azure Databricks and Event Hubs, where the associated documentation goes into more depth about this pattern.

Microsoft Azure