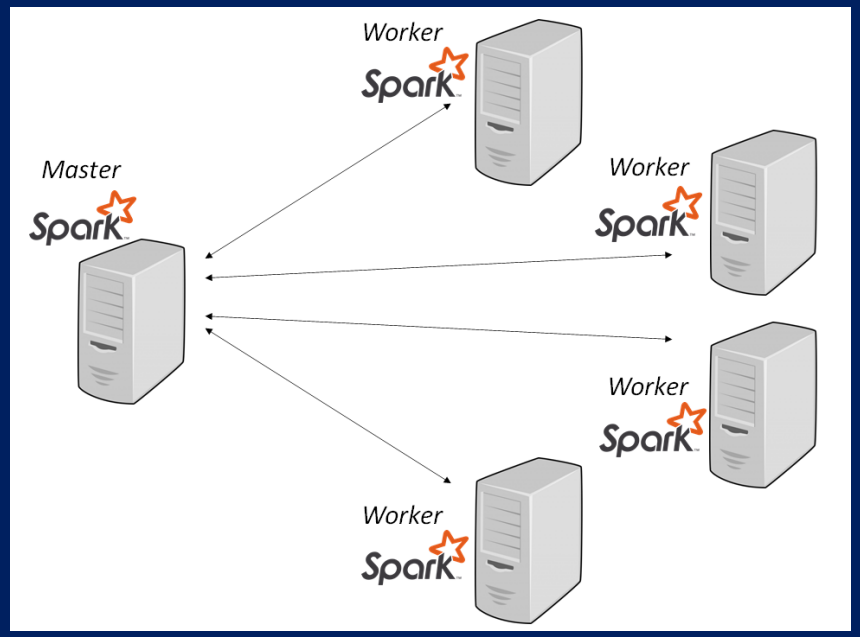
Databricks

Clusters



Bryan Cafferky
Big Data and AI Consultant

https://github.com/bcafferky/shared

Where Are We Heading?

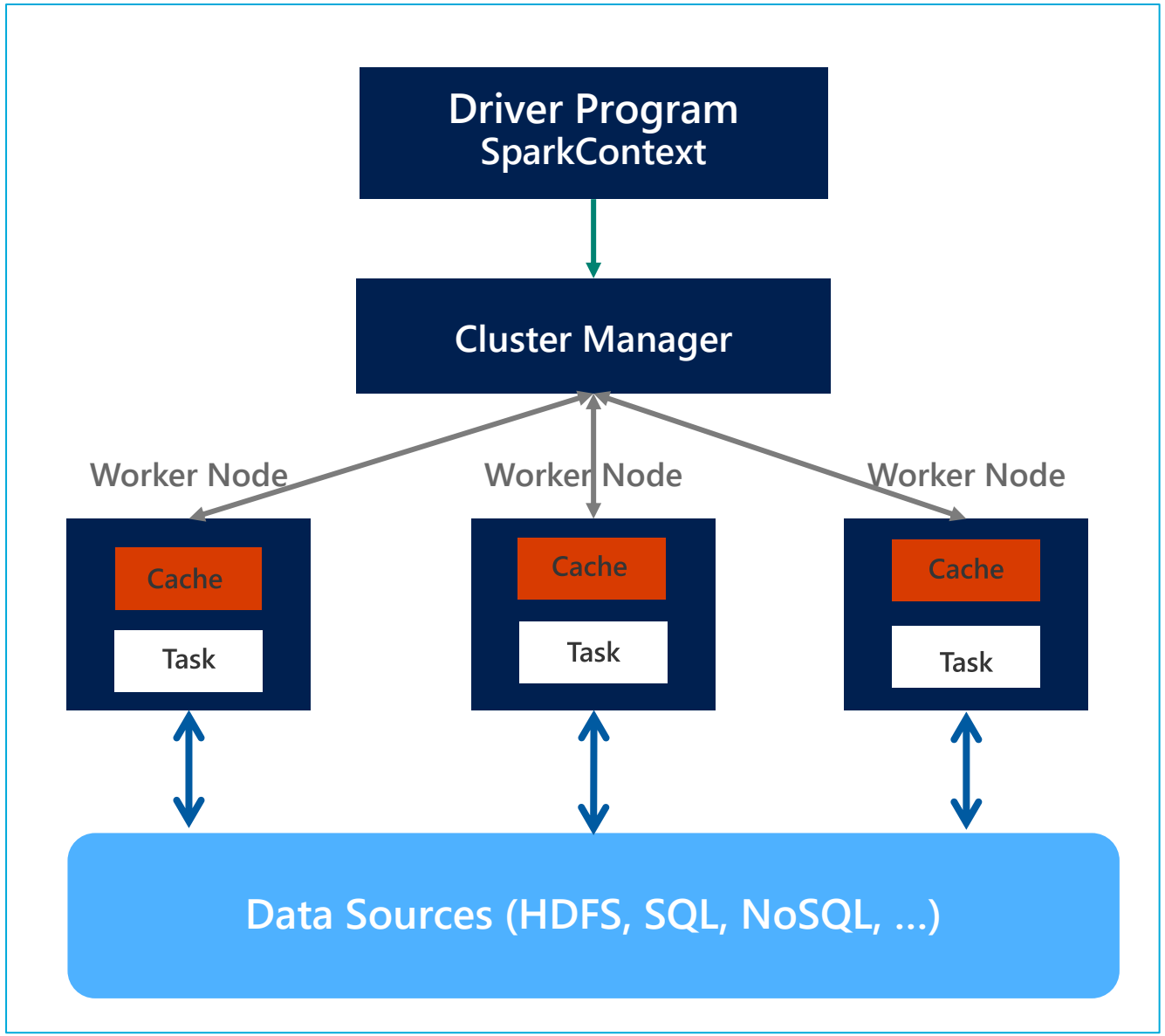
- What is a Databricks Cluster?
- Creating a Databricks Cluster
- Cluster Configuration Choices
- AdventureWorks Data Model
- Data Formats and Sources

What is a Databricks Cluster?

- An Apache Spark cluster with the Databricks driver and enhanced features.
- A set of computers to do Data Engineering workloads.
- All work is done on a cluster.
- The work is coordinated by a node called the Driver.
- The data is processed by Worker Nodes.
- A Single Node can have multiple parallel processes called Executors.

GENERAL SPARK CLUSTER ARCHITECTURE

- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).



Databricks Workspace

Microsoft Azure | PORTAL | brcaffer@microsoft.com

Azure Databricks | Home | Workspace | Recents | Data | Clusters | Jobs | Search

Demo_CognitiveServices_ImageTextExtraction (1) (Scala)

Detache | Workspace/Demo/Demo_CognitiveServices_ImageTextEx (1)

Blob Account Key : dsdsfddfdgf | Blob Account Name : bcafferkystoragexxx

Blob Container Name : bcafferkyc231

Cmd 1

Set Parameters to Notebook Variables

```
val blobcontainer = dbutils.widgets.get("containername")
val blobaccount = dbutils.widgets.get("accountname")
val blobaccountkey = dbutils.widgets.get("accountkey")
val bloburl = "wasbs://" + blobcontainer + "@" + blobaccount + ".blob.core.windows.net"
val blobconf = "fs.azure.account.key." + blobaccount + ".blob.core.windows.net"
```

Cmd 2

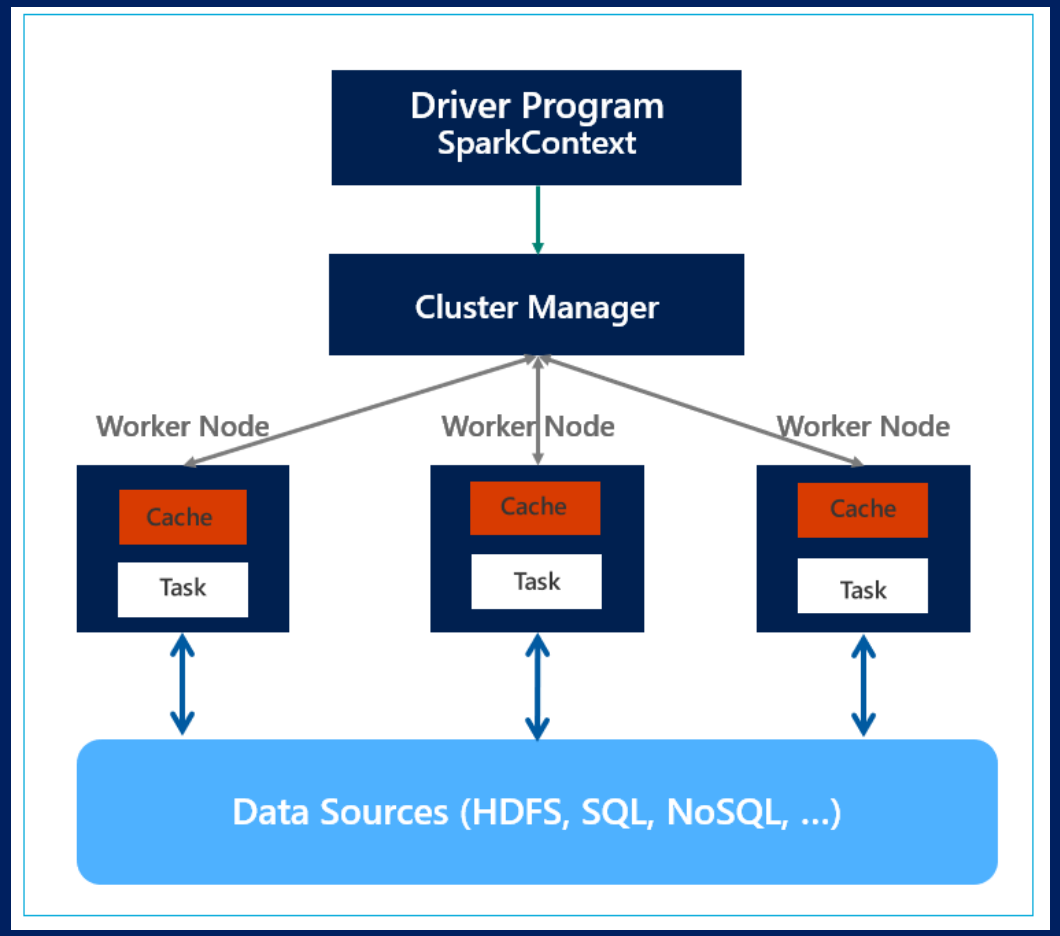
Mount the Blob Container to DBFS to /mnt/blob

```
dbutils.fs.mount(
  source = bloburl,
  mountPoint = "/mnt/blobimages",
  extraConfigs = Map(blobconf -> blobaccountkey))
```

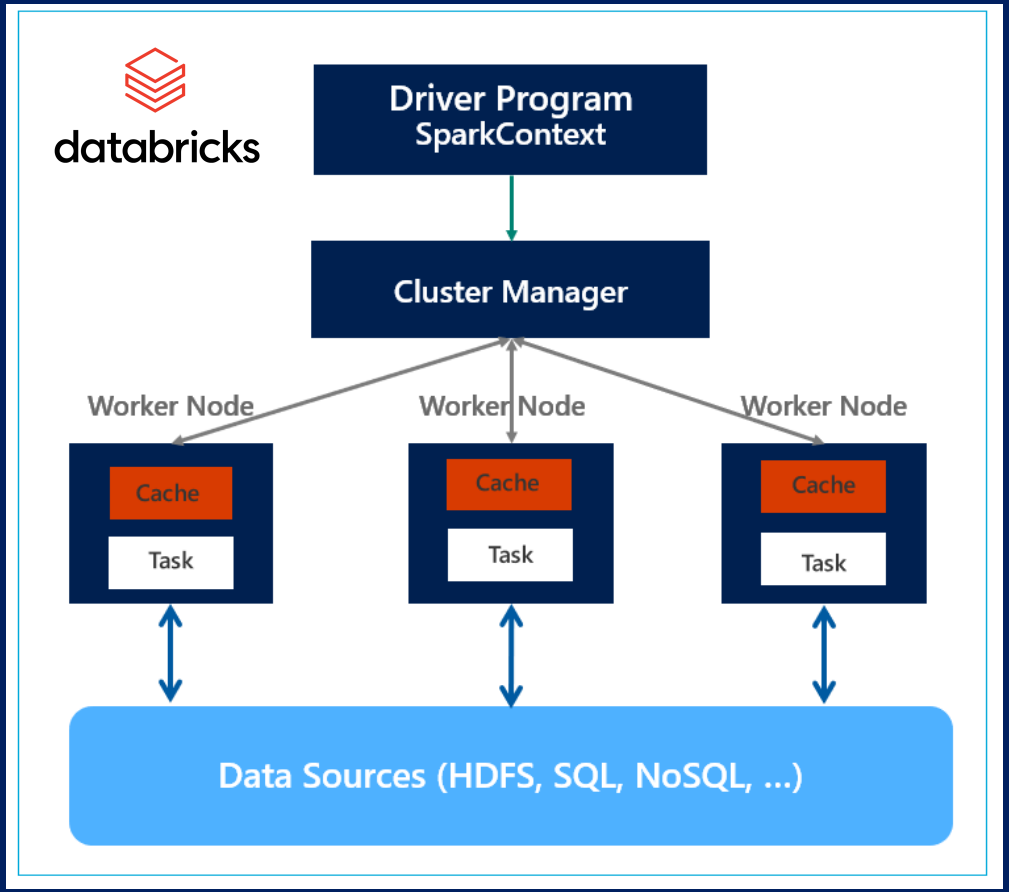
Cmd 3

View files in the DBFS mount point (which is mounted to Azure Storag

```
%fs ls /mnt/blobimages
```



Creating a Databricks Cluster



Creating a Databricks Cluster – Main Screen

Create Cluster

New Cluster | Cancel | Create Cluster | DBU / hour: 2.25 - 6.75 | 2-8 Workers:28-112 GB Memory, 8 1 Driver:14 GB Memory, 4 Cores

Cluster name

mycluster

Cluster mode

Standard

Databricks runtime version | Learn more

Runtime: 9.1 LTS (Scala 2.12, Spark 3.1.2)

50% promotional discount applied to Photon during preview

Autopilot options

Enable autoscaling

Terminate after 120 minutes of inactivity

Worker type | Min workers | Max workers

Standard_DS3_v2 14 GB Memory, 4 Cores | 2 | 8 | Spot instances

New Configure separate pools for workers and drivers for flexibility. Learn more

Driver type

Same as worker 14 GB Memory, 4 Cores

DBU / hour: 2.25 - 6.75 | Standard_DS3_v2

Advanced options

Databricks Cluster Type

Databricks Runtime

Let Databricks Scale Cluster as Needed

Pause the Cluster After X Minutes of No Use

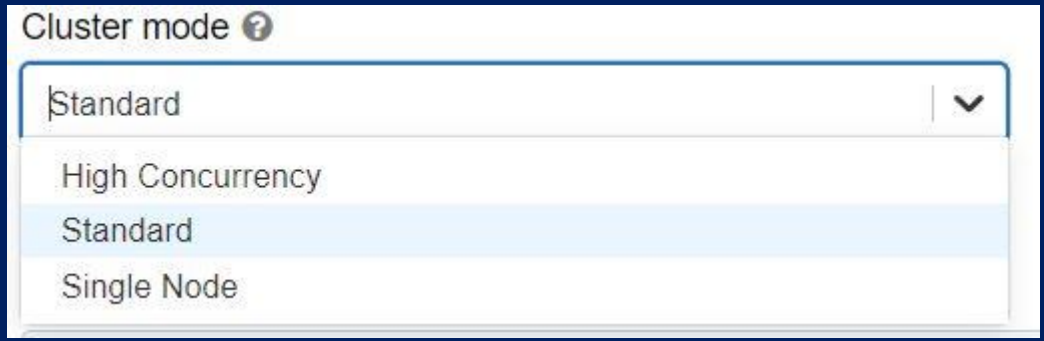
Minimum and Maximum # of Worker Nodes

Keeps VMs Ready to Use for Cluster

Driver Node VM Type

Customer Cluster Configuration

Creating a Databricks Cluster – Cluster Mode

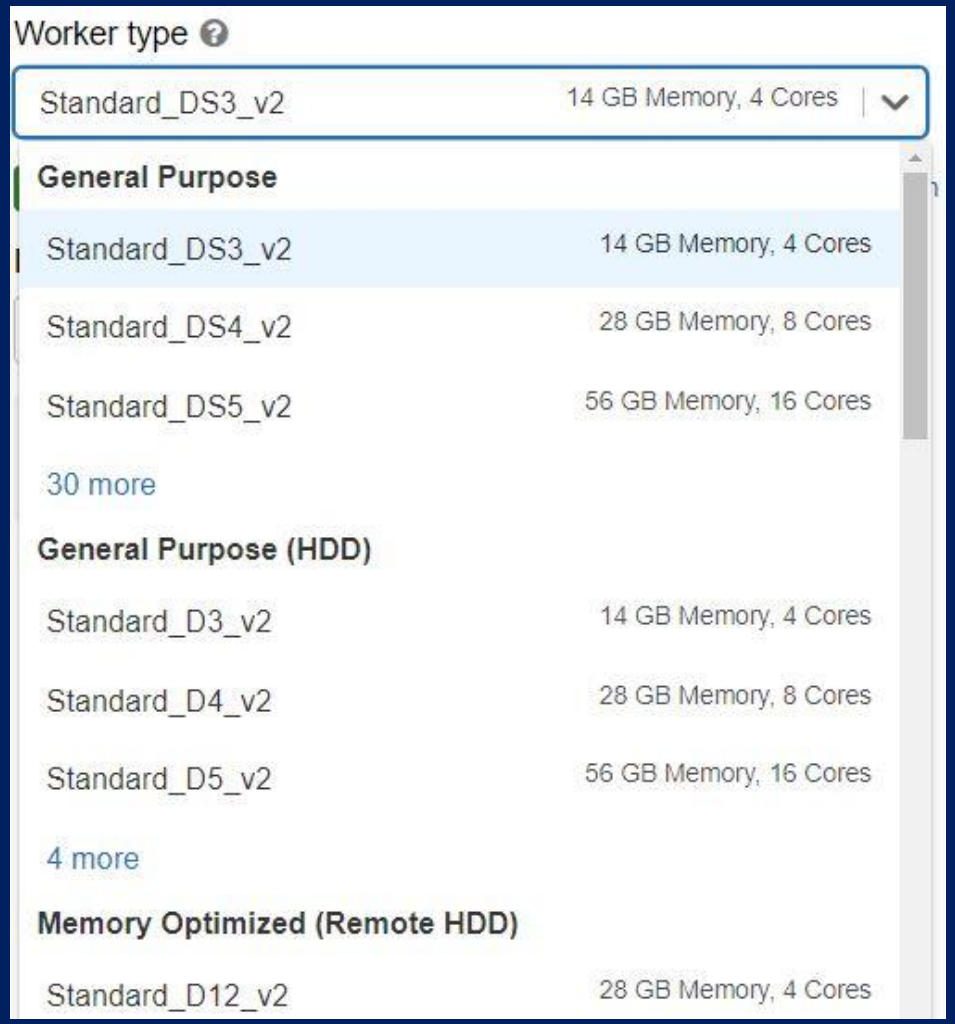


Option	Meaning
High Concurrency	Optimized for sharing the cluster with many concurrent users. Scala is not supported in this cluster mode.
Standard	Standard Spark Cluster mode of a driver and workers
Single Node	Just a Driver Node is created to minimize costs. Good for learning.

Creating a Databricks Cluster – Autopilot Options

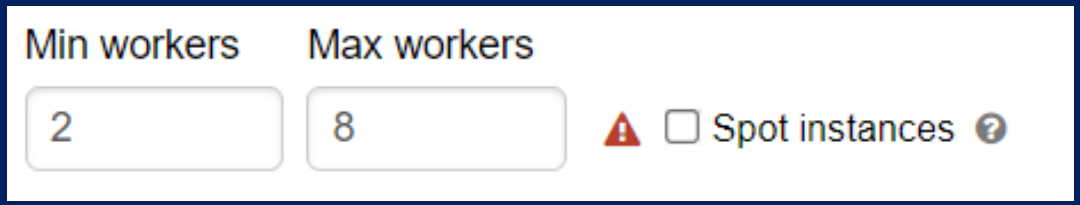
Option	Meaning
Enable autoscaling	Let Databricks automatically increase or decrease the number of worker nodes based on the workload demanded at any given time.
Terminate after	To save money, pause the cluster after the specificized number of minutes have passed with no usage.

Creating a Databricks Cluster – Worker Type



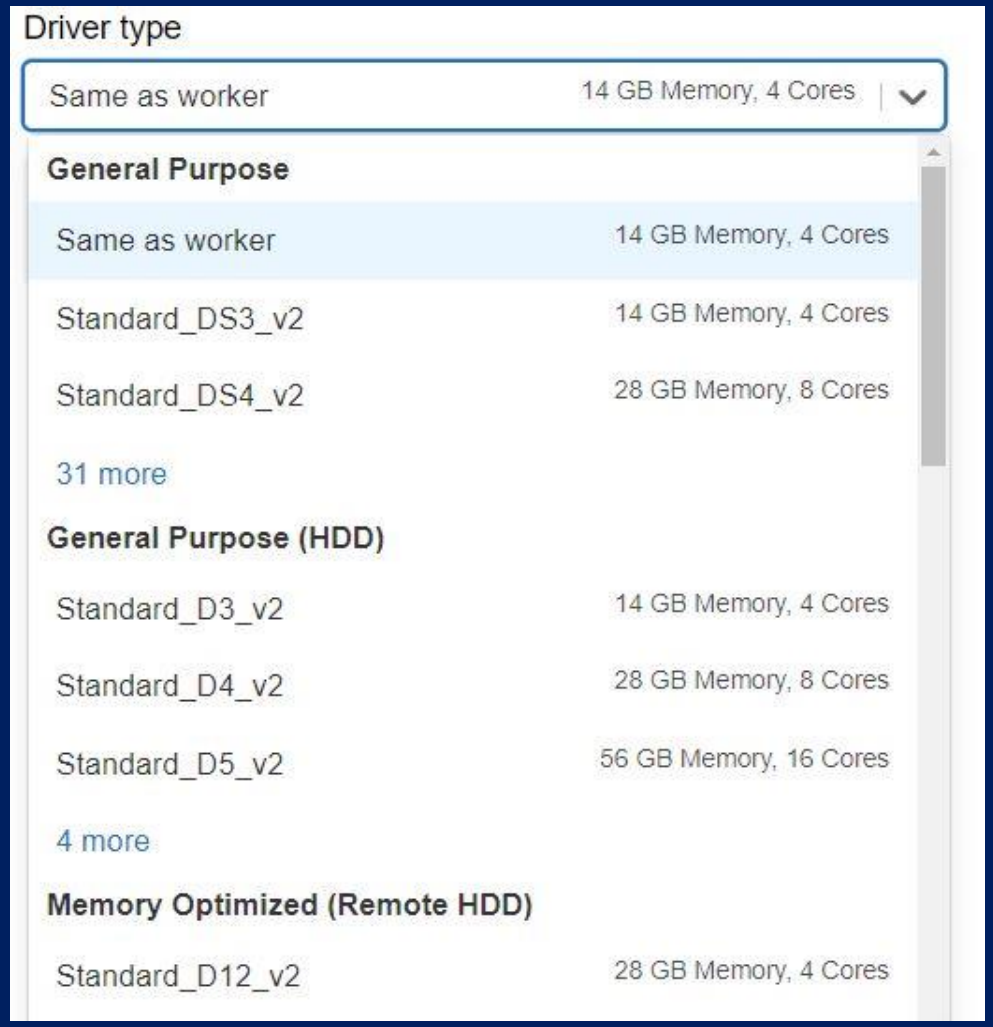
Option	Meaning
Worker type	Select the Virtual Machine Type to use for the worker nodes.

Creating a Databricks Cluster – Cluster Mode



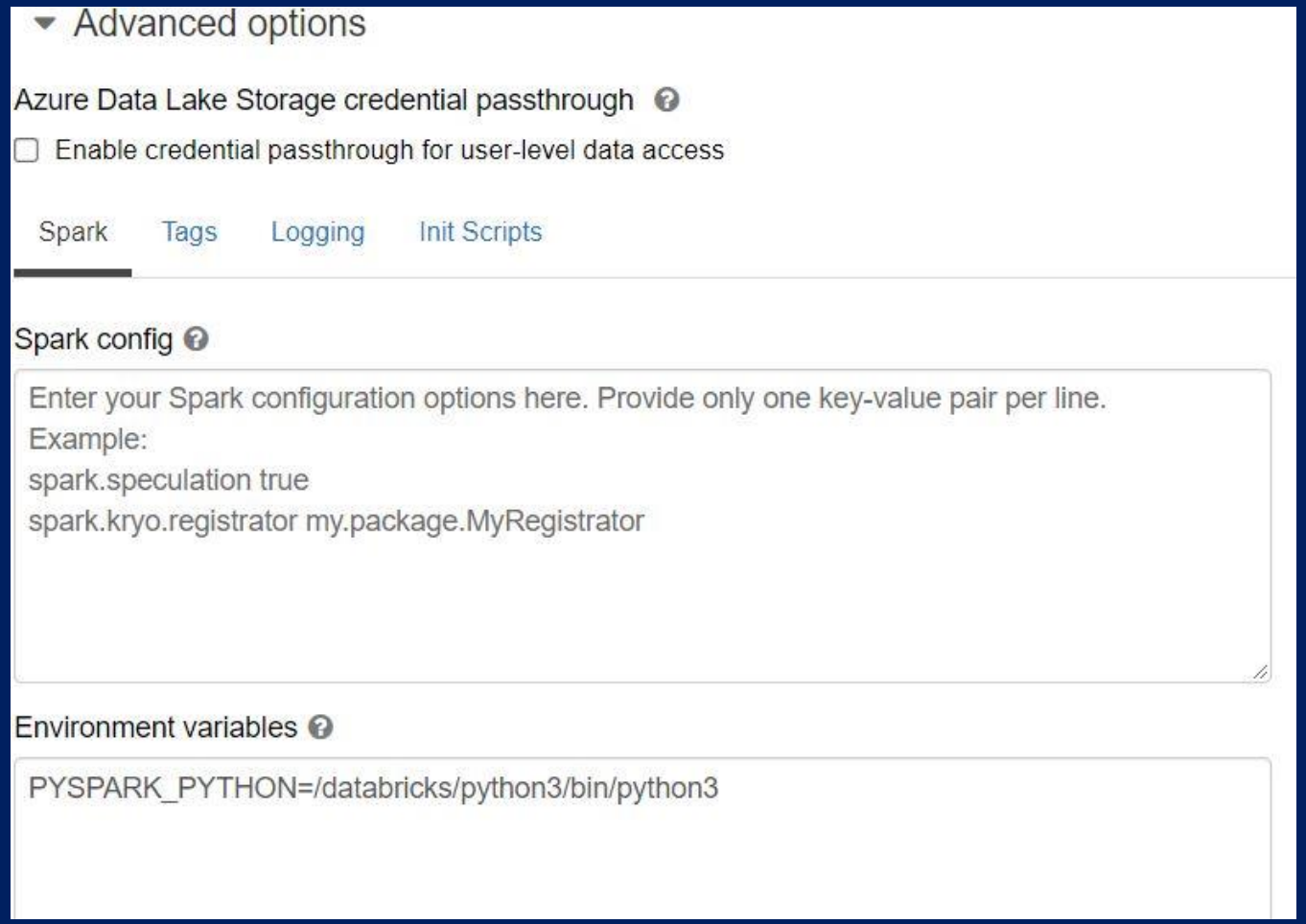
Option	Meaning
Min workers	The smallest number of worker nodes that must be maintained for the cluster.
Max workers	The largest number of worker nodes that may be created for the cluster.
Spot instances	Spot instances provide a way to save money by letting Databricks take unused VMs available in the cloud platform as discounted costs. https://techcommunity.microsoft.com/t5/analytics-on-azure-blog/azure-databricks-and-azure-spot-vms-save-cost-by-leveraging/ba-p/2374187

Creating a Databricks Cluster – Driver Type



Option	Meaning
Driver type	Select the Virtual Machine Type to use for the cluster driver. Sometimes you want this to be different like wanted to do more work on the driver.

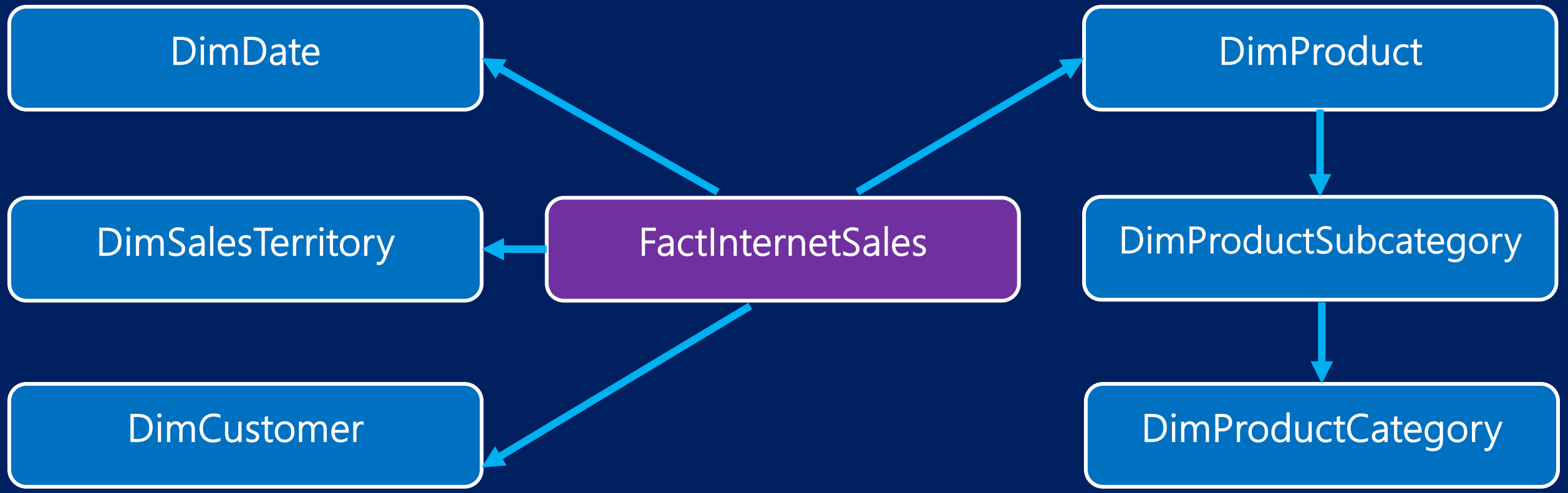
Creating a Databricks Cluster – Advanced options



Option	Meaning
Advanced options	Click on Advanced options to open the Advanced options form. Here you can customize the cluster configuration and perform additional set up work like installing open source libraries.

About the data

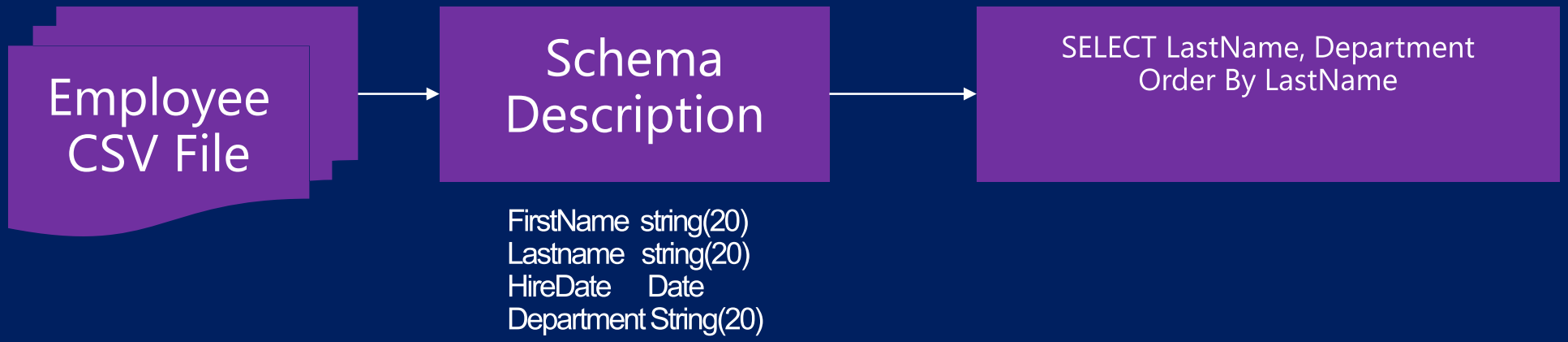
AdventureWorks Data Model



- Each table has a common key to join them.

Schema On Read

- Data Is Not in an RDMS
- External File Is Described Structurally



Built-In Databricks Datasets:
https://docs.databricks.com/data/databricks-datasets.html

Databricks Supported Data Types and Sources

Text Files

- CSV
- JSON
- XML
- TXT

Public Cloud

- Databases
- NoSQL
- Cloud Storage
- Data Services

Big Data Formats

- Parquet
- Delta Lake
- ORC
- Avro

Where We've Been

- What is a Databricks Cluster?
- Creating a Databricks Cluster
- Cluster Configuration Choices
- AdventureWorks Data Model
- Data Formats and Sources