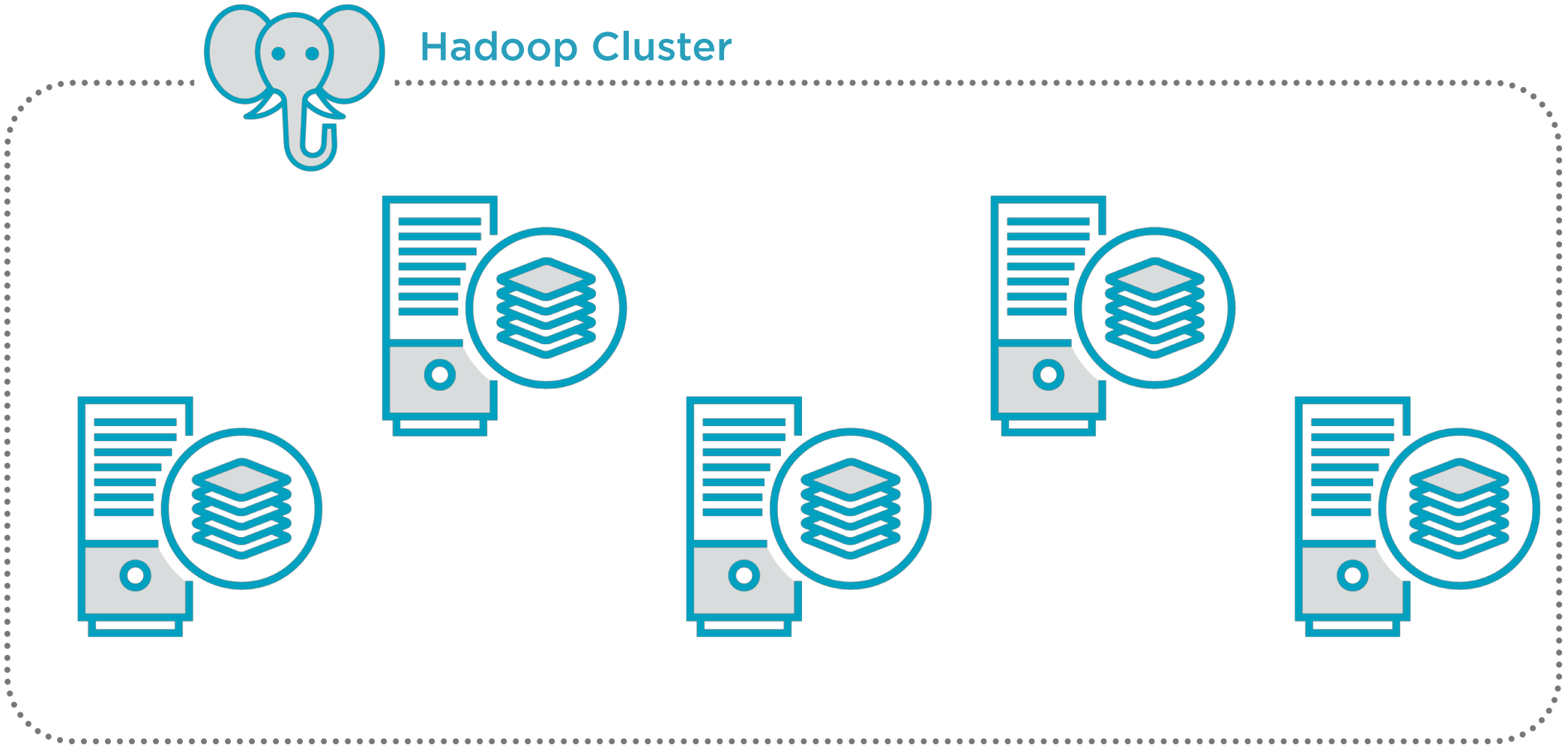
Using a Data Lake Store as External Repository with a Hadoop Cluster

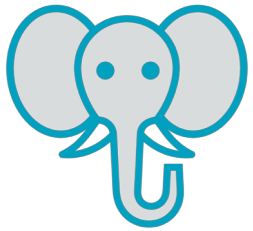
Xavier Morera

PASSIONATE ABOUT ENTERPRISE SEARCH AND BIG DATA

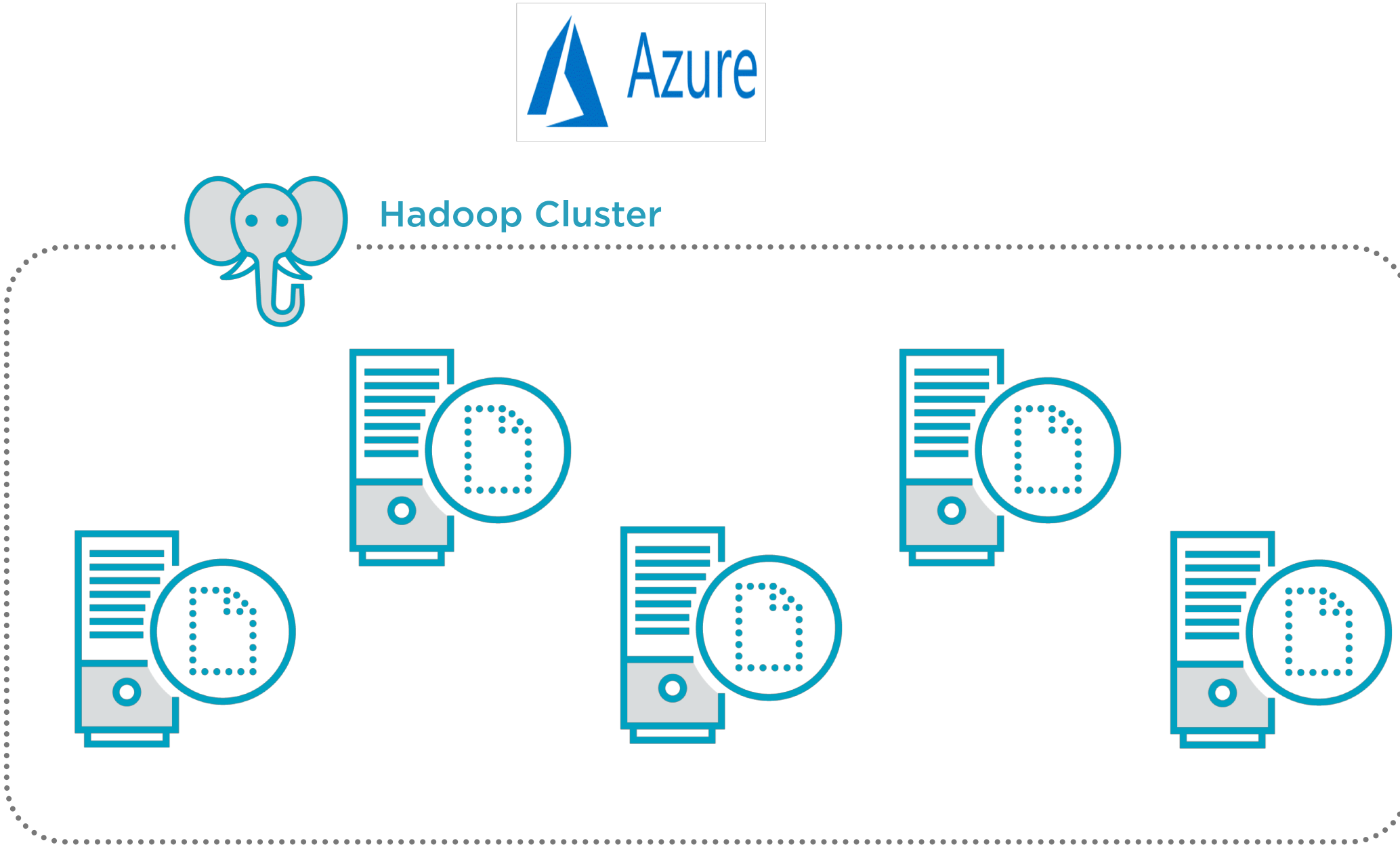
@xmorera www.xaviermorera.com

Hadoop Distributed File System

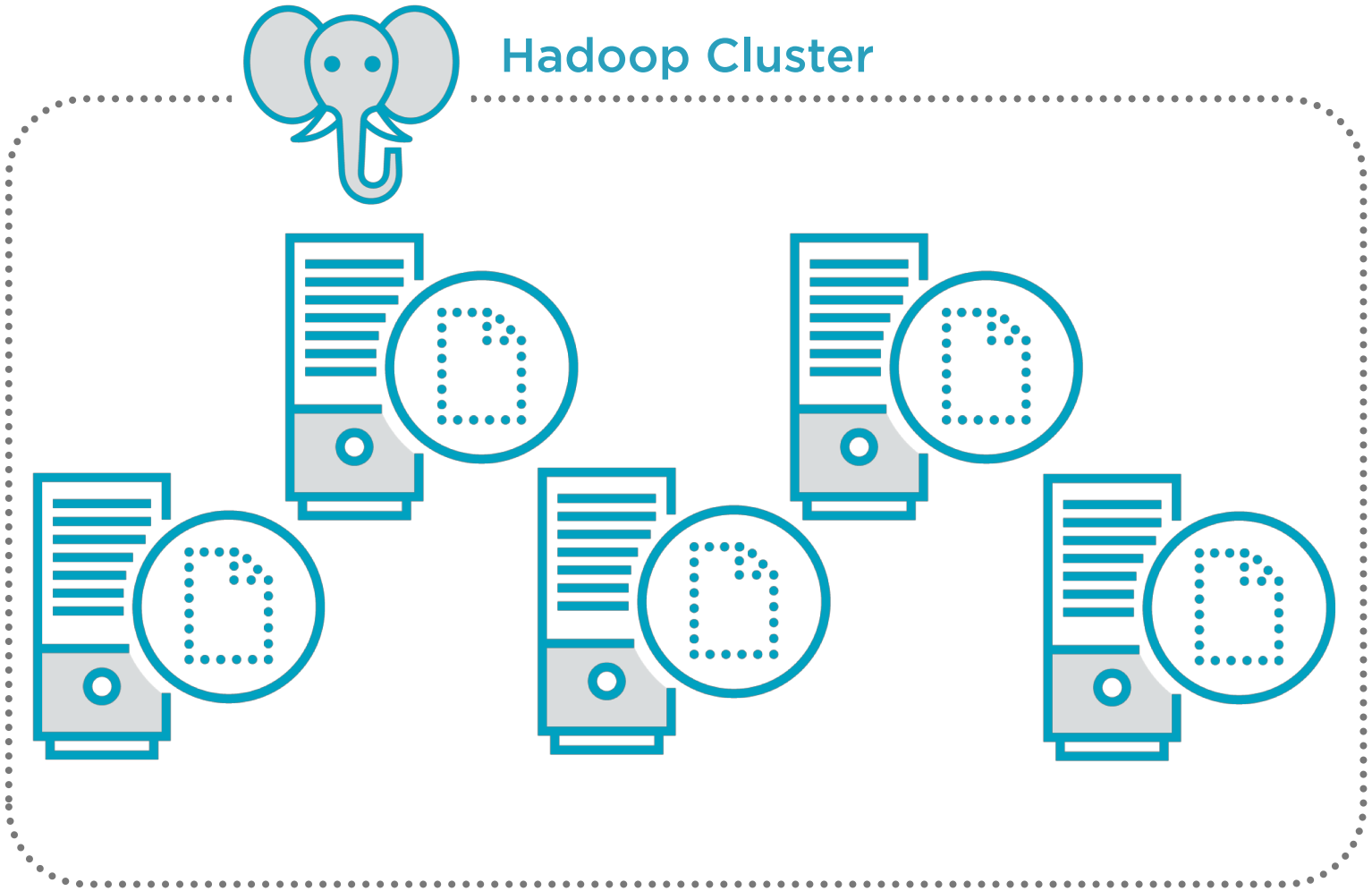




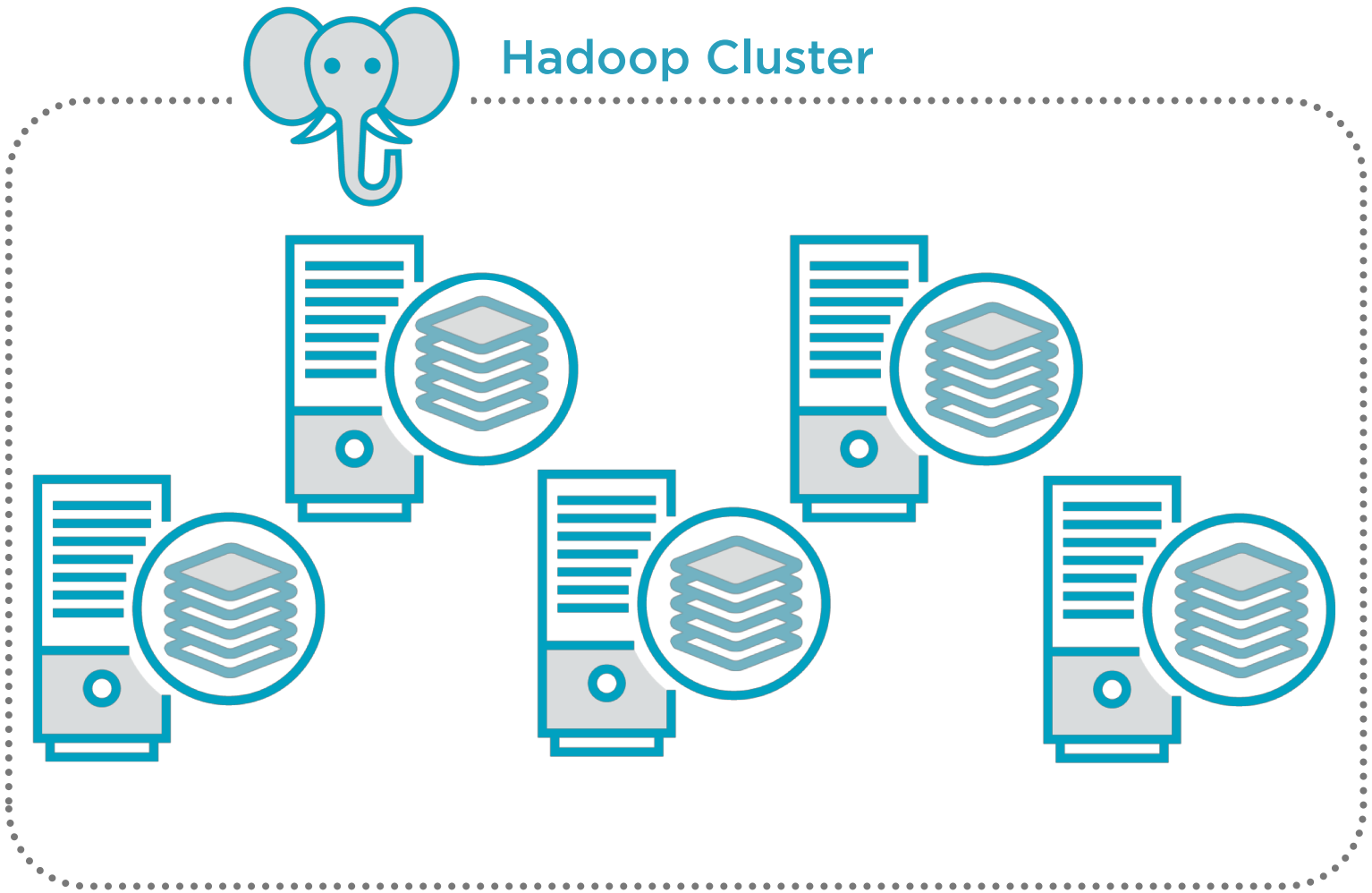
Hadoop Cluster


Azure
Hadoop Cluster

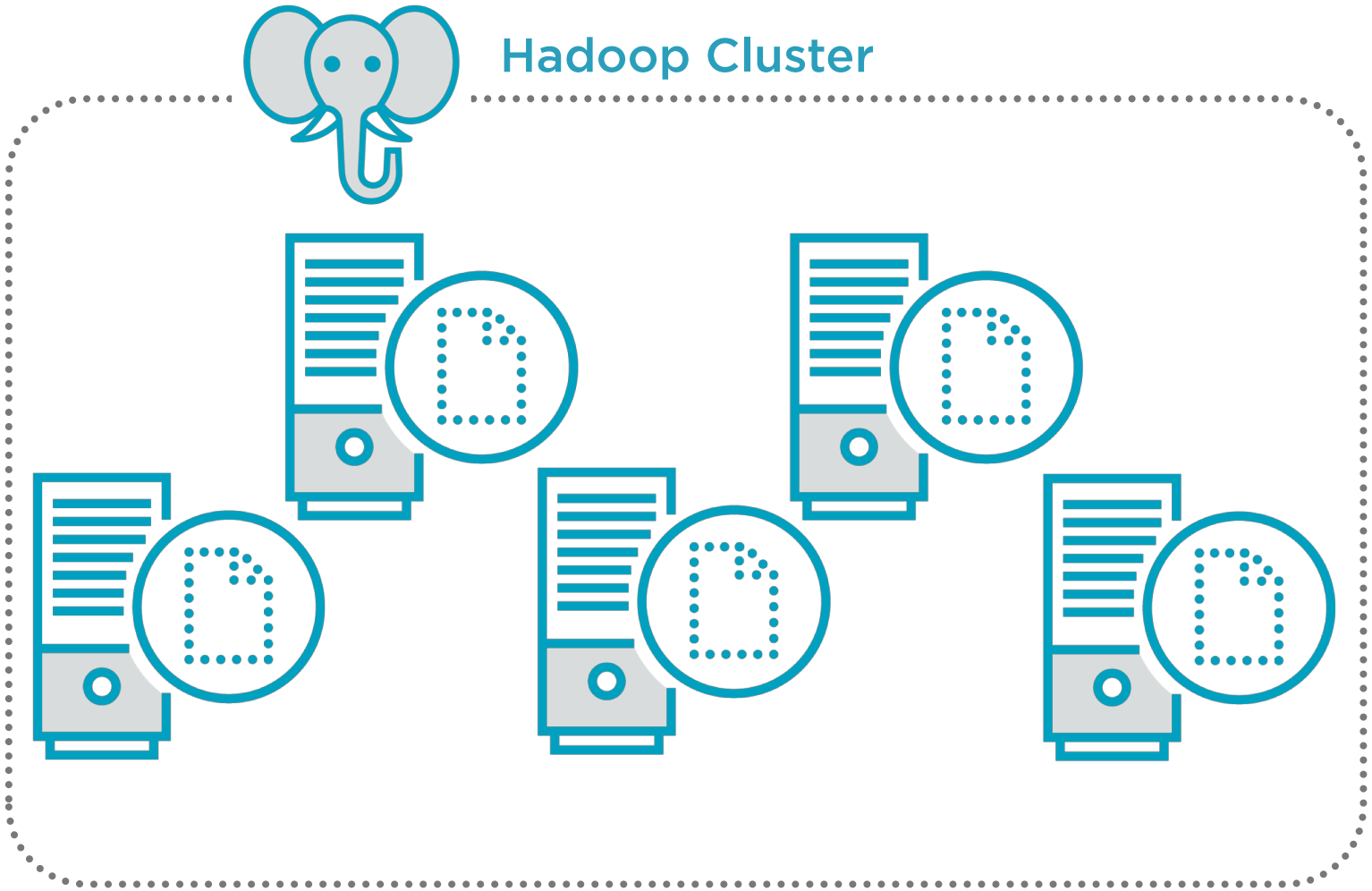
data

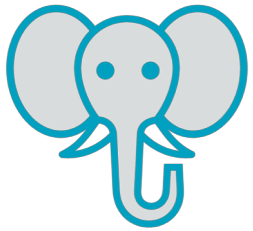


data

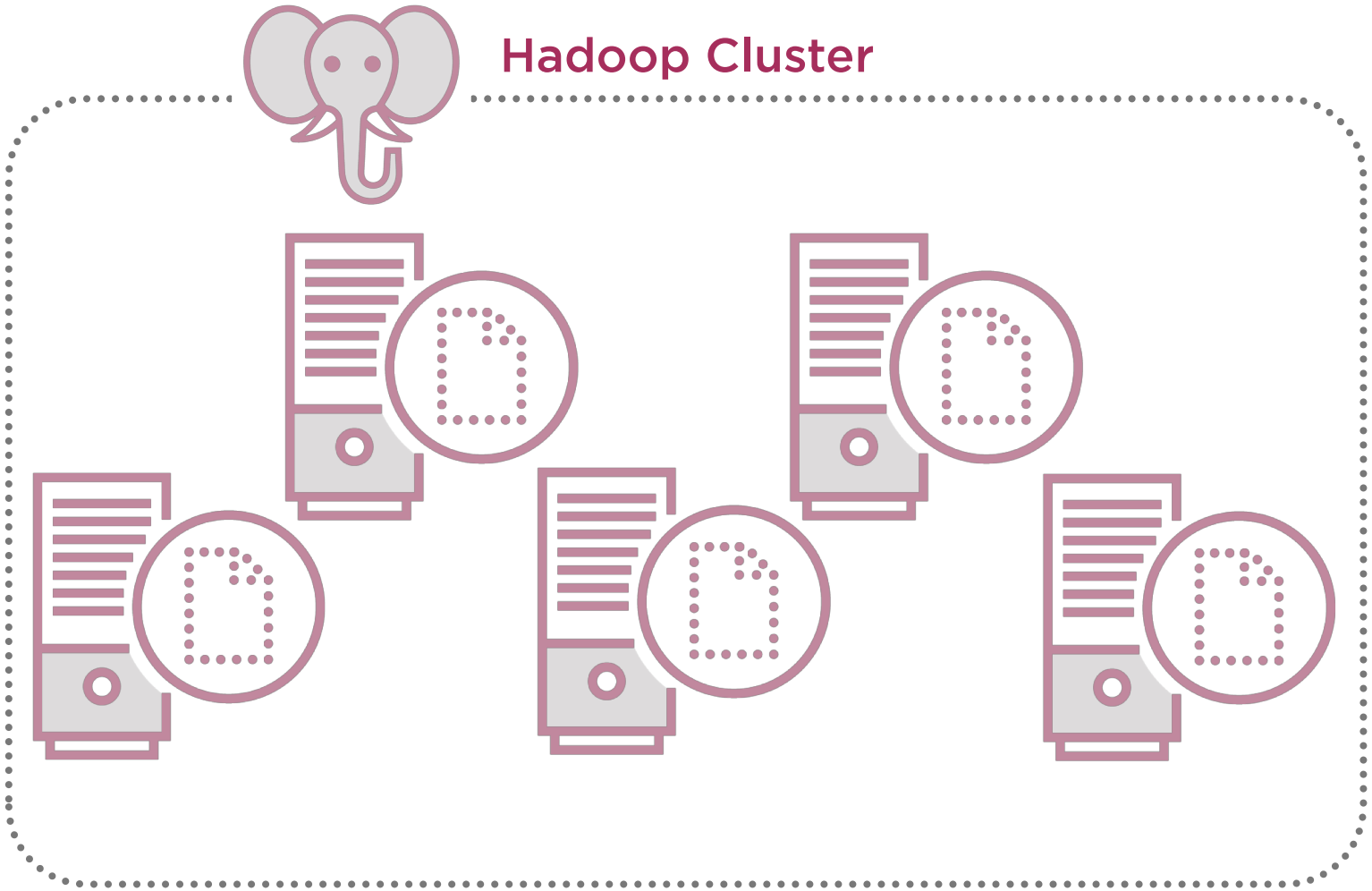


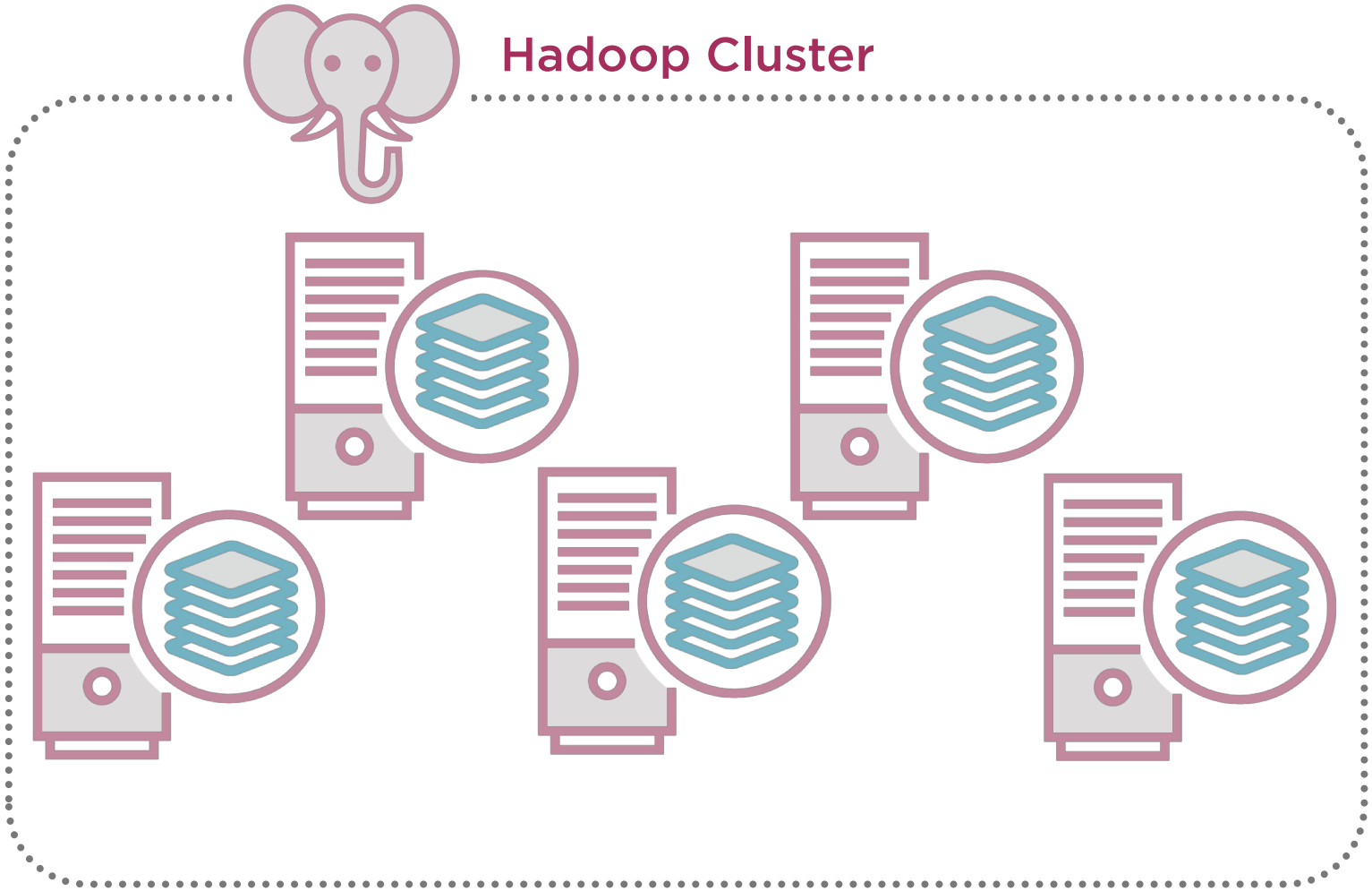
data





Hadoop Cluster


Azure
Hadoop Cluster


Azure
Hadoop Cluster

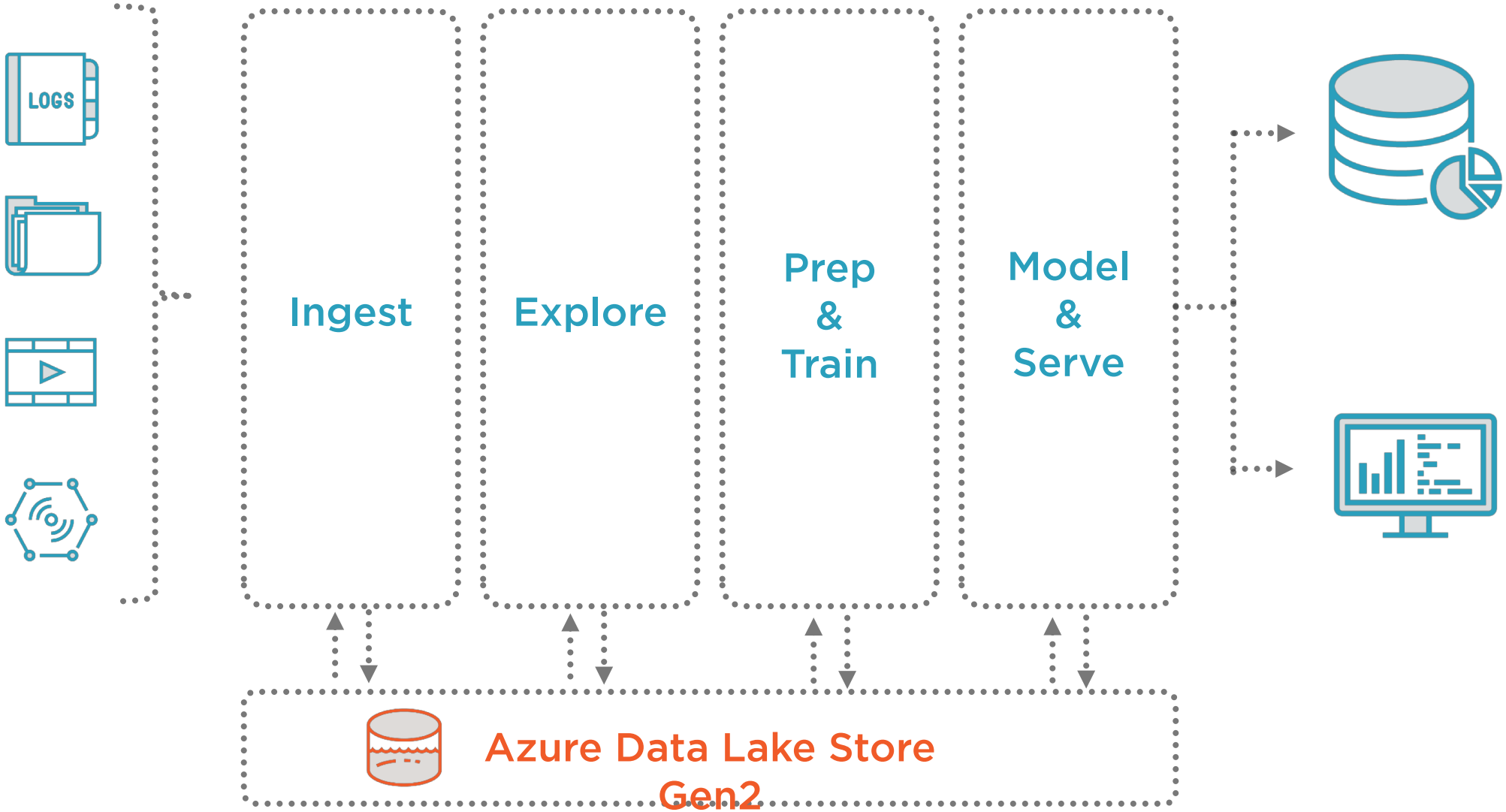
End to End Analytics

Modern data warehouse

Real-time analytics

Advanced analytics

End to End Analytics



Demo

Setting up a Data Lake Store Gen2 Capable HDInsight Cluster

Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
Storage accounts
Virtual machines
Virtual networks
Network security groups
Azure Active Directory
App Services
SQL databases
Data Lake Storage Gen1
Data Lake Analytics
Azure Cosmos DB
Load balancers
Security Center
Cost Management + Bill...
Help + support

HDInsight
by Microsoft

Quick create | Custom (size, settings, apps)

1 Basics
Configure basic settings

2 Storage
Set storage settings

3 Summary
Confirm configurations

This cluster may take up to 20 minutes to create.

Basics

* Cluster type
Spark 2.3 (HDI 3.6)

* Cluster login username
admin

* Cluster login password
••••••••••••

Secure Shell (SSH) username
sshuser

Use same password as cluster login

* Resource group
datalake-ps
Create new

* Location
East US 2

Click here to view cores usage.

Next

Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
Storage accounts
Virtual machines
Virtual networks
Network security groups
Azure Active Directory
App Services
SQL databases
Data Lake Storage Gen1
Data Lake Analytics
Azure Cosmos DB
Load balancers
Security Center
Cost Management + Bill...
Help + support

datalakegen2hdi

HDInsight cluster

Search (Ctrl+/)
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Tools

Settings
Cluster size
Quota limits
SSH + Cluster login
Data Lake Storage Gen1
Storage accounts
Applications
Script actions
External metastores
HDInsight partner

Move Delete Refresh

Resource group (change)
datalake-ps

Status
Running

Location
West US 2

Subscription (change)
Visual Studio Enterprise

Subscription ID
251d961b-545e-4b26-abdd-adaa7d6b5568

Tags (change)
Click here to add tags

Learn more
Documentation

Cluster type, HDI version
Spark 2.3 (HDI 3.6)

URL
https://datalakegen2hdi.azurehdinsight.net

Getting started
Quickstart

Cluster dashboards

Cluster management interfaces

Ambari home
Ambari views
Zeppelin notebook
Jupyter notebook
Spark history server
Yarn

Cluster size

```
sshuser@datalakegen2hdi-ssh.azurehdinsight.net's password:
Welcome to Ubuntu 16.04.5 LTS (GNU/Linux 4.15.0-1023-azure x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.

New release '18.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.



Welcome to Spark on HDInsight.

Last login: Wed Nov  7 18:45:51 2018 from 190.7.211.138
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

sshuser@hn0-datala:~$
```

Demo

Running Spark Jobs with Data Stored in ADLS Gen2

```
SLF4J: Found binding in [jar:file:/usr/hdp/2.6.5.3003-25/spark2/jars/slf4j-log4j12-1.7.16
.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/2.6.5.3003-25/spark_llap/spark-llap-assembly-1
.0.0.2.6.5.3003-25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLeve
l).
Spark context Web UI available at http://hn0-datala.cnon3am5j2durnzpw0k3ruekfh.xx.interna
l.cloudapp.net:4040
Spark context available as 'sc' (master = yarn, app id = application_1541616092236_0004).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.0.2.6.5.3003-25
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

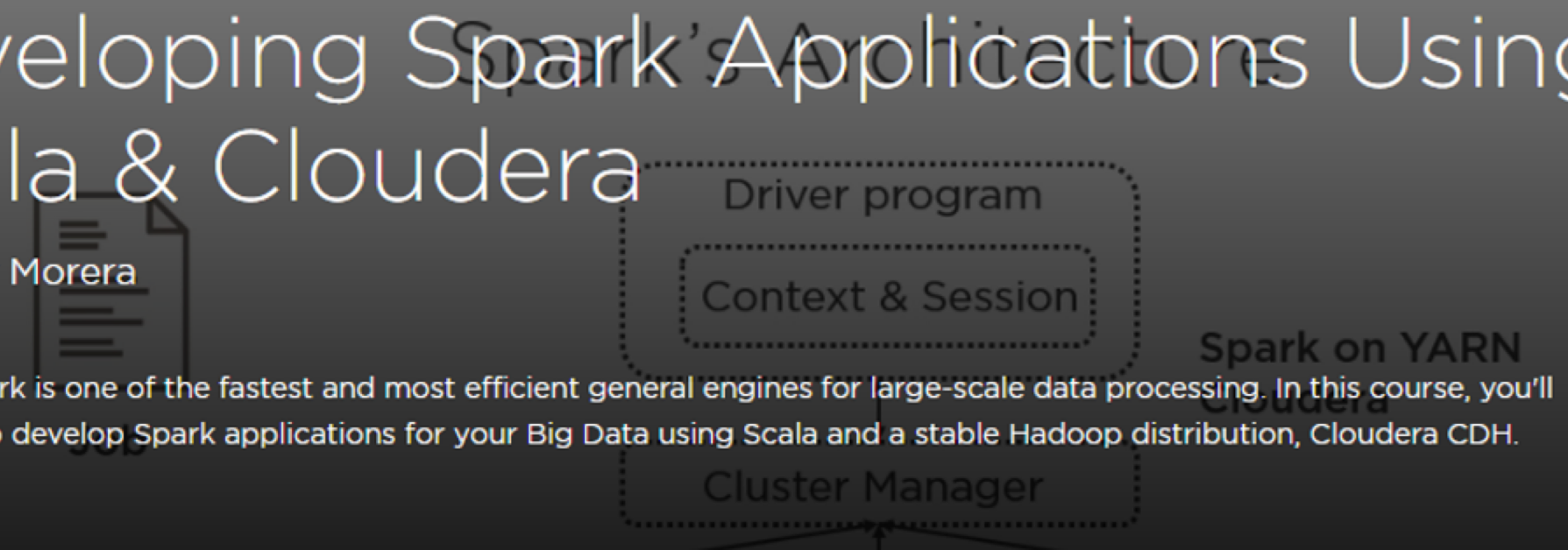
Developing Spark Applications with Python & Cloudera

by Xavier Morera

Apache Spark is one of the fastest and most efficient general engines for large-scale data processing. In this course, you will learn how to develop Spark applications for your Big Data using Python and a stable Hadoop distribution, Cloudera CDH.

Developing Spark Applications Using Scala & Cloudera

by Xavier Morera



Apache Spark is one of the fastest and most efficient general engines for large-scale data processing. In this course, you'll learn how to develop Spark applications for your Big Data using Scala and a stable Hadoop distribution, Cloudera CDH.

jupyter Untitled Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help PySpark3

Code CellToolbar

In [2]:
```
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
```

In [*]:
```
posts_df = spark.read.option('header', 'True').csv('abfs://datalakegen2hdifs@datalakesaps.dfs.core.windows.net/data/Pos
posts_df.select('Id', 'Score', 'Title').show()
```

```
+---+-----+--------------------+
| Id|Score|               Title|
+---+-----+--------------------+
|  4|  506|While applying op...|
|  6|  223|Percentage width ...|
|  9| 1546| Calculate age in C#|
| 11| 1205|Calculate relativ...|
| 13|  495|Determine a User'...|
| 14|  318|Difference betwee...|
| 16|   94|Filling a DataSet...|
| 17|  137|Binary Data in MySQL|
| 19|  246|What is the faste...|
| 24|  117|Throw an error in...|
| 25|  113|How to use the C ...|
| 34|   61|Unloading a ByteA...|
| 36|  112|Check for changes...|
| 39|   65|Reliable timer in...|
| 42|  234|Best way to allow...|
| 48|  215|Multiple submit b...|
| 59|   70|How do I get a di...|
| 61|   33|Office 2007 File ...|
| 66|   58|Paging a collecti...|
| 72|   30|How do I add exis...|
+---+-----+--------------------+
only showing top 20 rows
```

Takeaway

HDFS is a Hadoop Cluster's Data Lake

Cluster deprovisioned means no more data

External repository for transient clusters

Azure Data Lake Store comes into play

- Gen1 (HDFS based)
- Gen2 (Blob Storage based)

Takeaway

Spin up clusters as needed

Run Big Data jobs

- Spark
- Hive, Impala...

Deprovision when no longer needed

Never lose your data

- Possible to store metadata

Takeaway

End to End Analytics

- Modern data warehouse
- Real-time analytics
- Advanced analytics