

# **Introduction to .NET for Apache Spark**

Luis Quintanilla

[@ljquintanilla](#)

# Welcome



ML.NET Content Developer, Microsoft



[lqdev@outlook.com](mailto:lqdev@outlook.com)



<http://luisquintanilla.me>



[@ljquintanilla](https://twitter.com/ljquintanilla)



<https://github.com/lqdev>

# Code & Slides

<http://bit.ly/NYCCodeCamp2019>

# Agenda

- 01** What is Big Data?
- 02** Apache Spark
- 03** .NET for Apache Spark
- 04** Working with Data
- 05** Build Apps with .NET for Apache Spark

# What is **Big Data**?

# 3 V's of Big Data

## Volume

- Terabytes
- Petabytes

## Variety

- Structured
- Unstructured

## Velocity

- Batch
- Streaming

# Big Data Software



# What is **Apache Spark**?



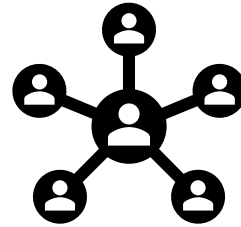
# Apache Spark



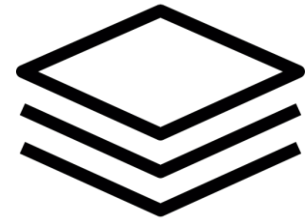
Open  
Source



Cross  
Platform

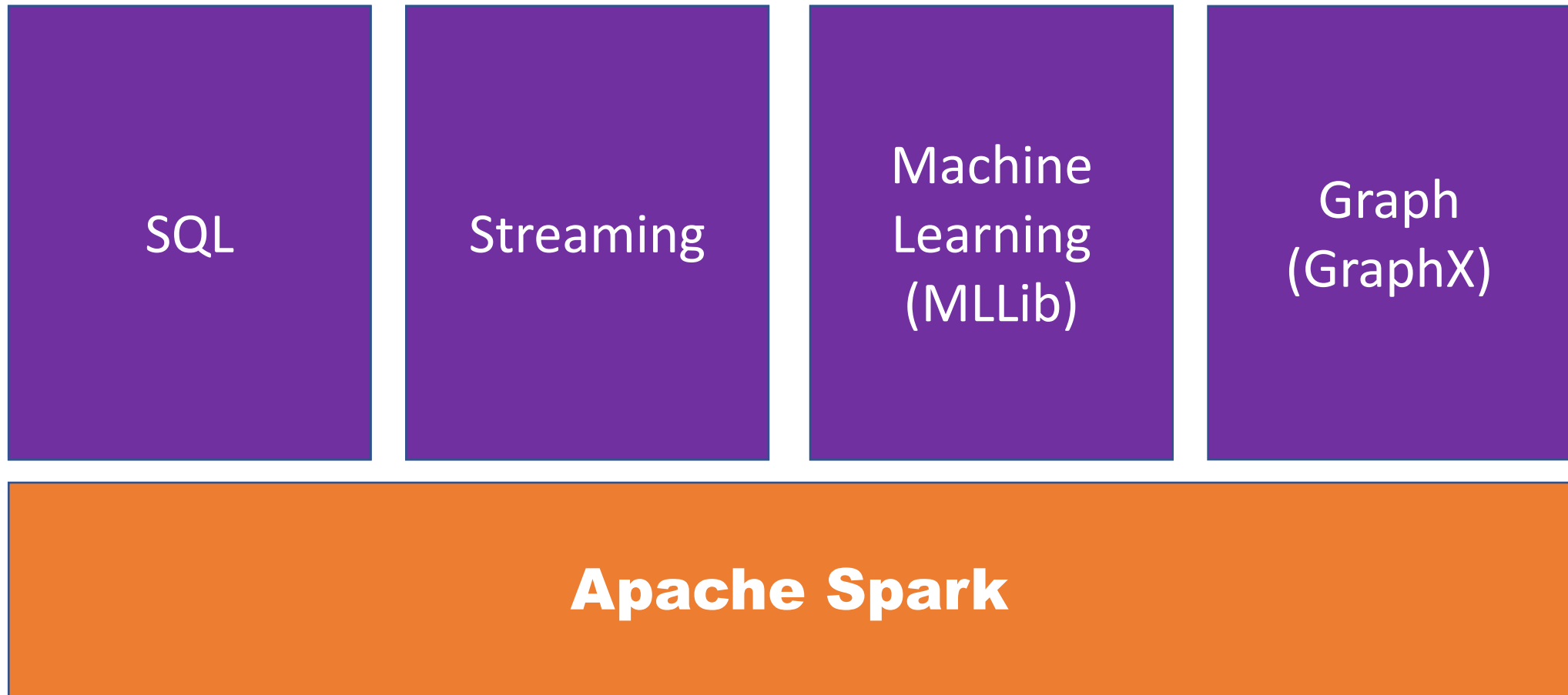


Distributed\*

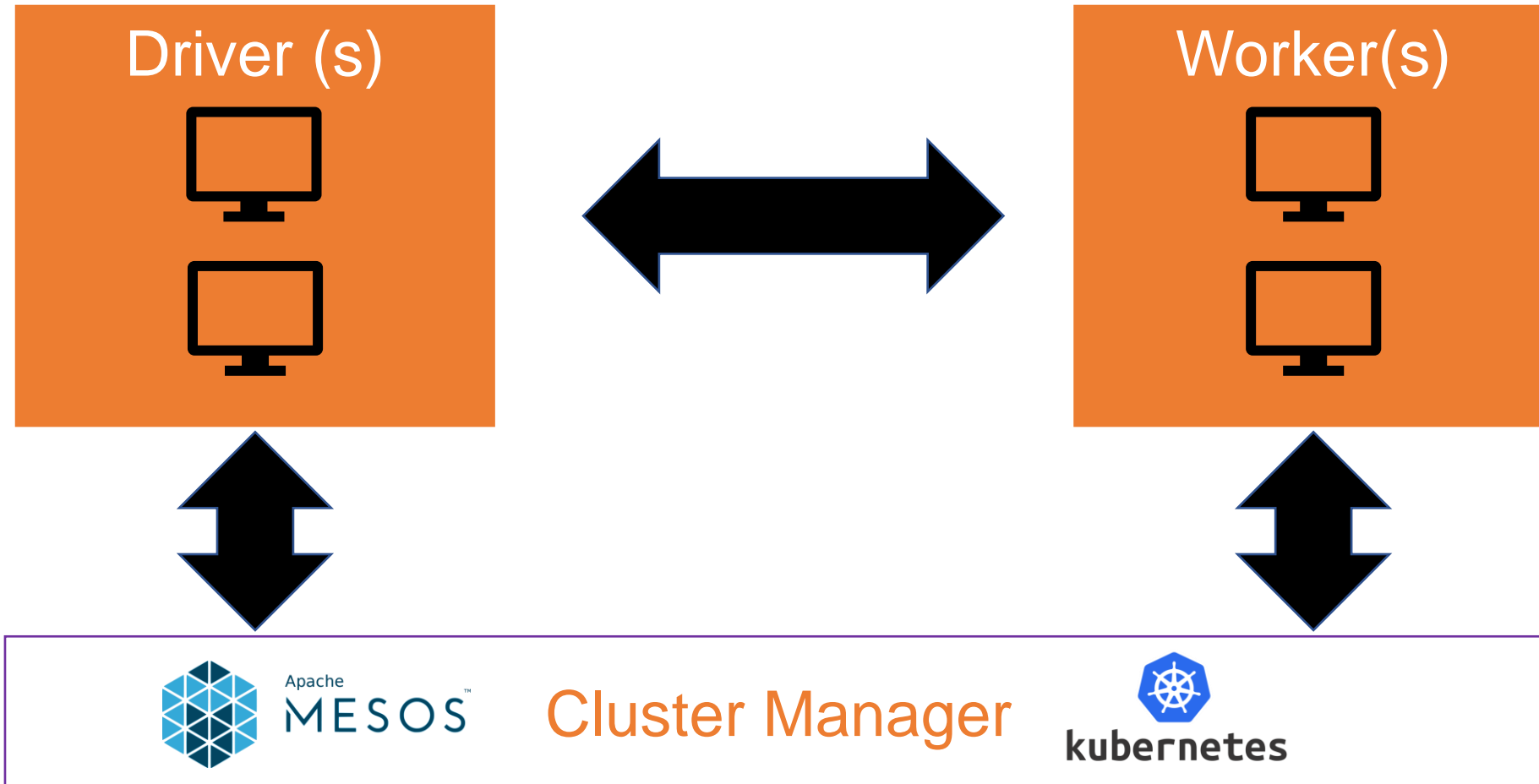


Extensible

# Apache Spark Architecture



# Apache Spark Execution



# Interacting with Apache Spark

```
b.MutableMetricsFactory).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Using Spark's repl log4j profile: org/apache/spark/log4j-defaults-repl.properties
To adjust logging level use sc.setLogLevel("INFO")
Welcome to

  ____  _
 / ___|| | | |
| |___| |_| |
|___ \  __| |
   __| | | |
  |___|_|_|_|

 version 1.6.1

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_91)
Type in expressions to have them evaluated.
Type :help for more information.
█
```

# Shell

The image shows a web browser window with multiple tabs. The active tab is titled 'localhost8888/notesbook/Spark!!: jupyter'. The address bar shows the URL. Below the browser window, there is a Jupyter interface. The top of the Jupyter interface has a menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', and 'Help'. Below the menu bar is a toolbar with various icons for file operations and code execution. The main area of the Jupyter interface displays a notebook with a single code cell. The code cell contains the following Python code:

```
In [1]: import os
import sys

spark_home = os.environ.get('SPARK_HOME', None)
if not spark_home:
    raise ValueError("'SPARK_HOME' environment variable is not set")
sys.path.insert(0, os.path.join(spark_home, 'python'))
sys.path.insert(0, os.path.join(spark_home, 'python/lib/py4j-0.8.1-src.zip'))
sys.path.insert(0, os.path.join(spark_home, 'C:/spark-1.6.0-bin-hadoop2.6/python/lib/py4j-0.9-src.zip'))
execfile(os.path.join(spark_home, 'python/pyspark/shell.py'))
```

The output of the code cell shows the Spark logo, which is a stylized 'S' made of triangles. Below the logo, it says 'version 1.6.0'. Below that, it says 'Welcome to' followed by 'Using Python version 2.7.10 (default, Oct 21 2015 17:08:47)' and 'SparkContext available as sc, HiveContext available as sqlContext.' The bottom of the Jupyter interface shows the prompt 'In [ ]: |'.

# Notebooks

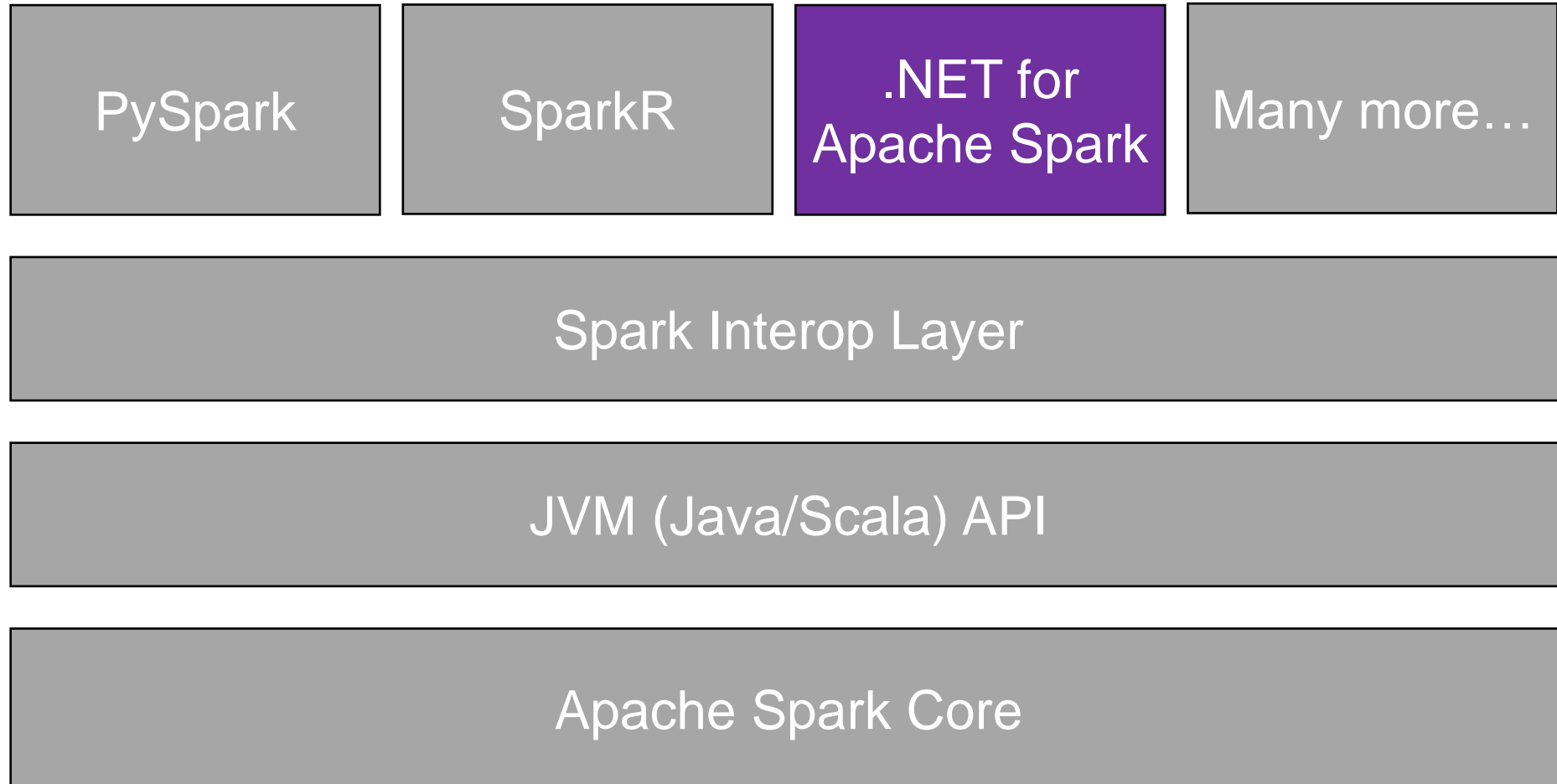
@ljquintanilla

```
// Create SparkSession
var sc =
    SparkSession
        .Builder()
        .AppName("Restaurant_Inspections_ETL")
        .GetOrCreate();

// Load data
DataFrame df =
    sc
        .Read()
        .Option("header", "true")
        .Option("inferSchema", "true")
        .Csv("Data/NYC-Restaurant-Inspections.csv");
```

# Traditional Apps

# Extending Apache Spark



# **.NET for Apache Spark**

# .NET for Apache Spark



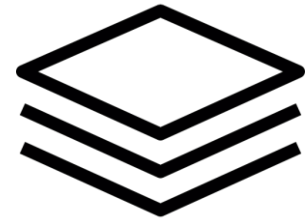
Open  
Source



Cross  
Platform

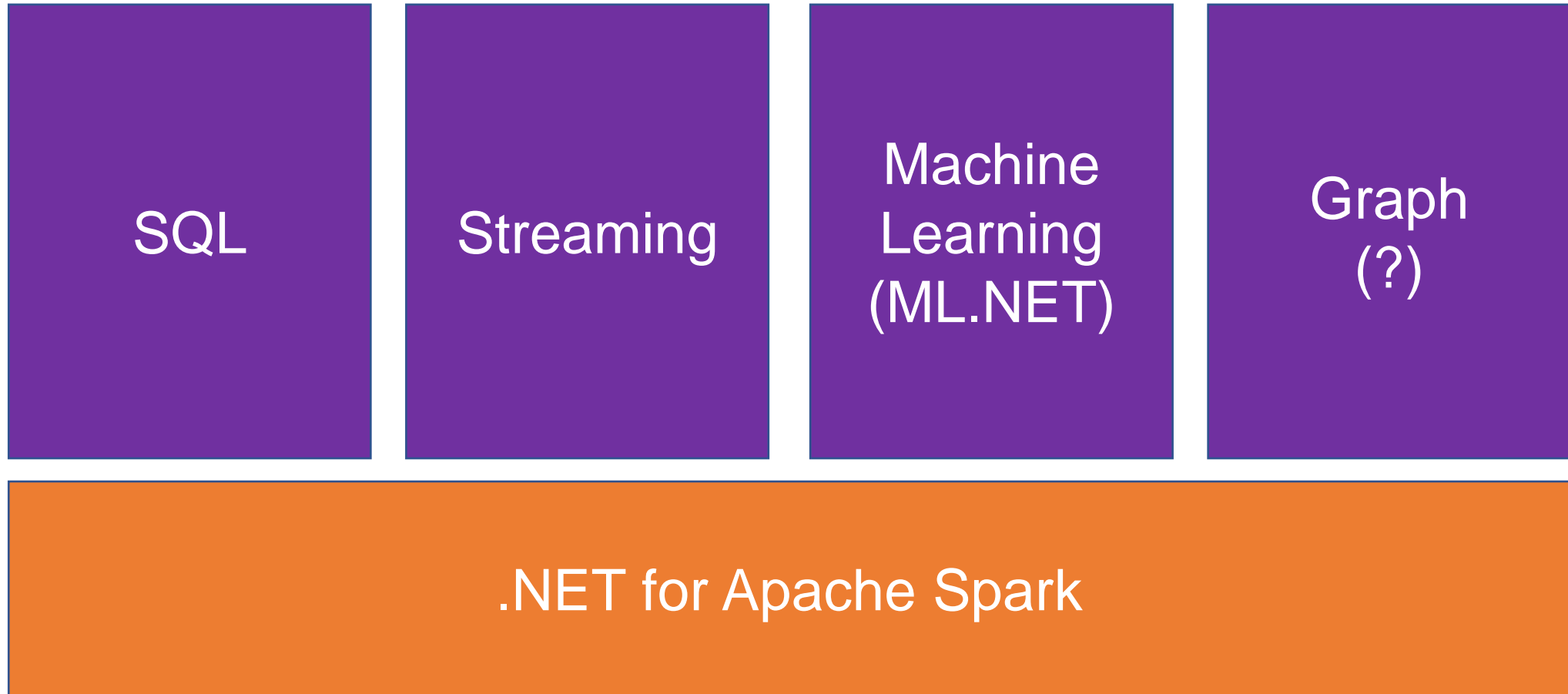


.NET  
Standard



Extensible

# .NET for Apache Spark Architecture





# ML.NET Framework

## Transforms

- Missing Values
- Feature Selection
- Normalization

## Trainers

- SVM
- K-Means
- Boosted Trees
- Logistic Regression

## Misc

- Data Loaders
- Evaluators

## Extensions

- TensorFlow
- ONNX

# Data in Apache Spark

Resilient  
Distributed  
Datasets (RDD)

DataFrames

Datasets

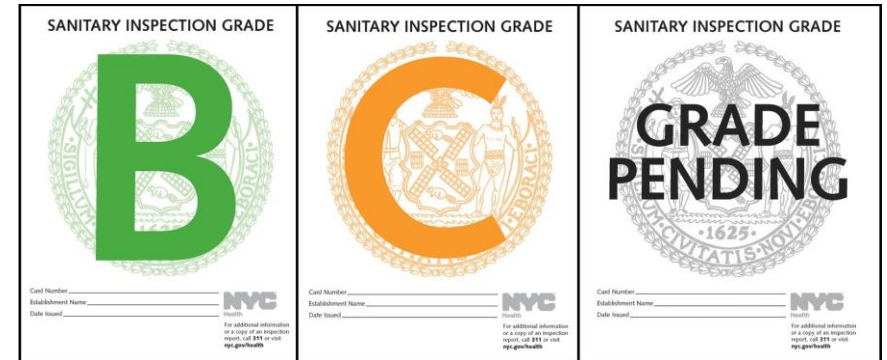
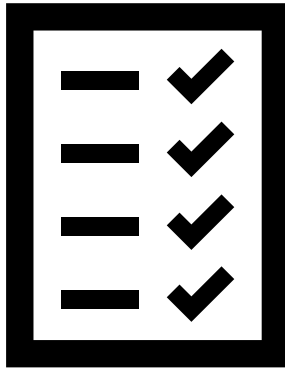
# **Demo**

# **.NET for Apache Spark**

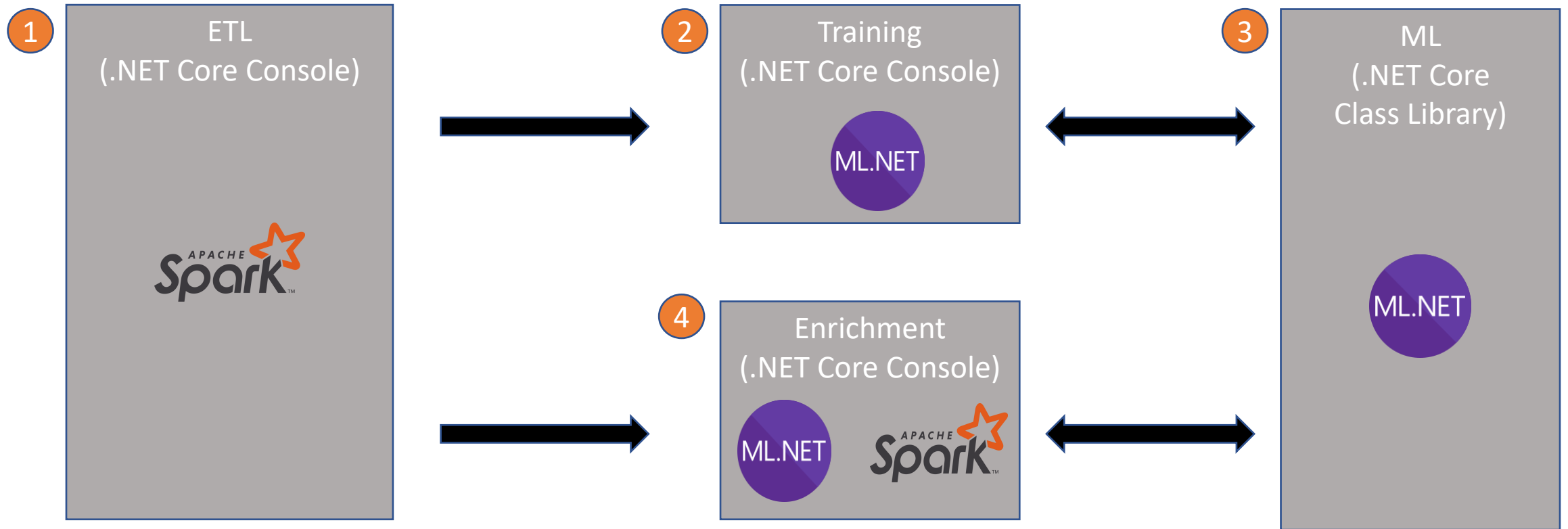
# **DataFrames**

# **Building Applications with .NET for Apache Spark**

# Restaurant Inspections



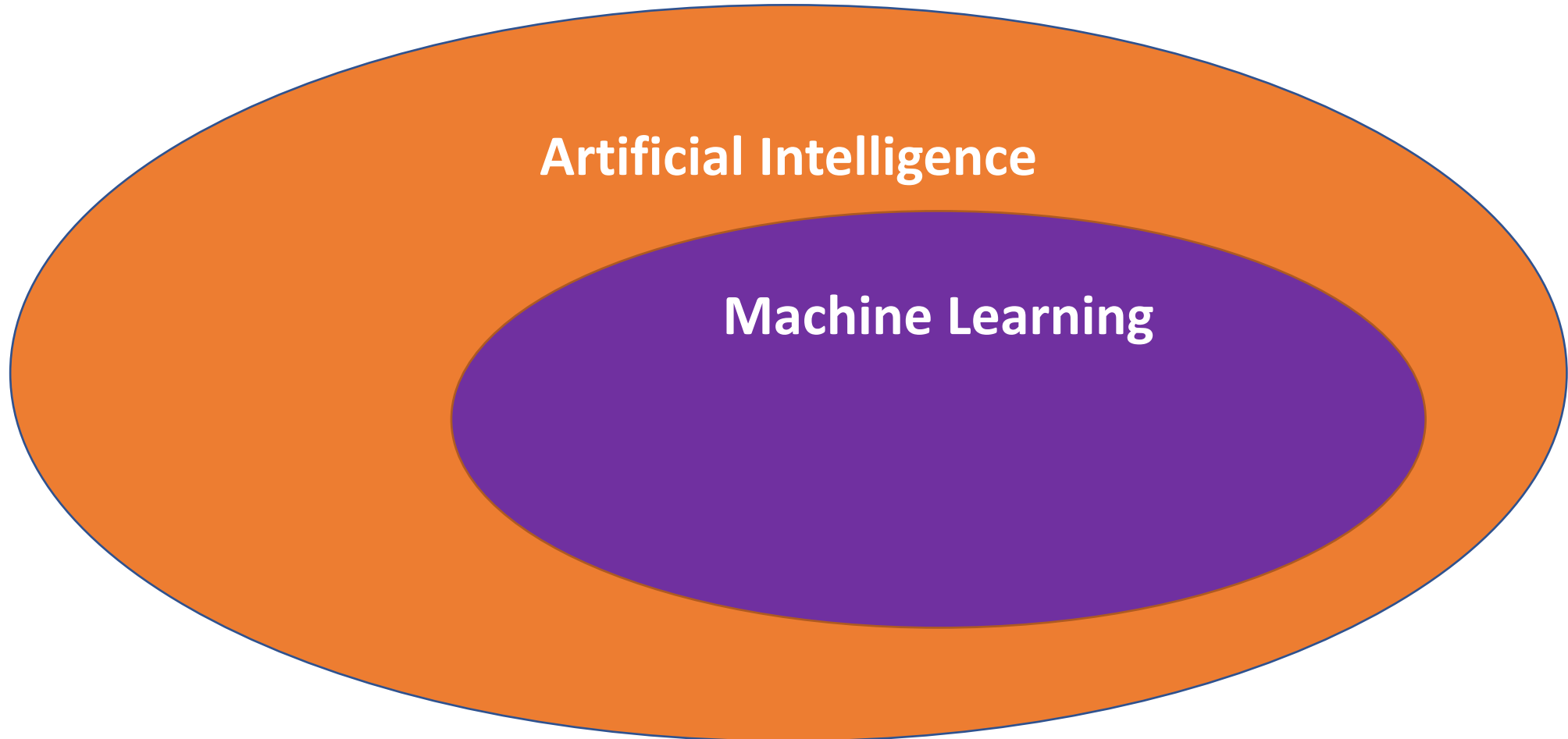
# Restaurant Inspections Architecture



# **Demo**

# **Restaurant Inspections ETL**

# What is Machine Learning?





# Machine Learning Tasks

## Supervised Learning

Regression

---

What is  
the price  
of a home  
in NYC?

Classification

---

Is this a  
dog or  
cat?

## Unsupervised Learning

Clustering

---

Customer  
segments  
in a  
database

# Classification Example

## Training Data

Species	Is Independent	Class
Canine	False	Dog
Feline	True	Cat
Feline	True	Cat
Canine	False	Dog
Canine	True	Dog

Features  
(input)

Label  
(output)

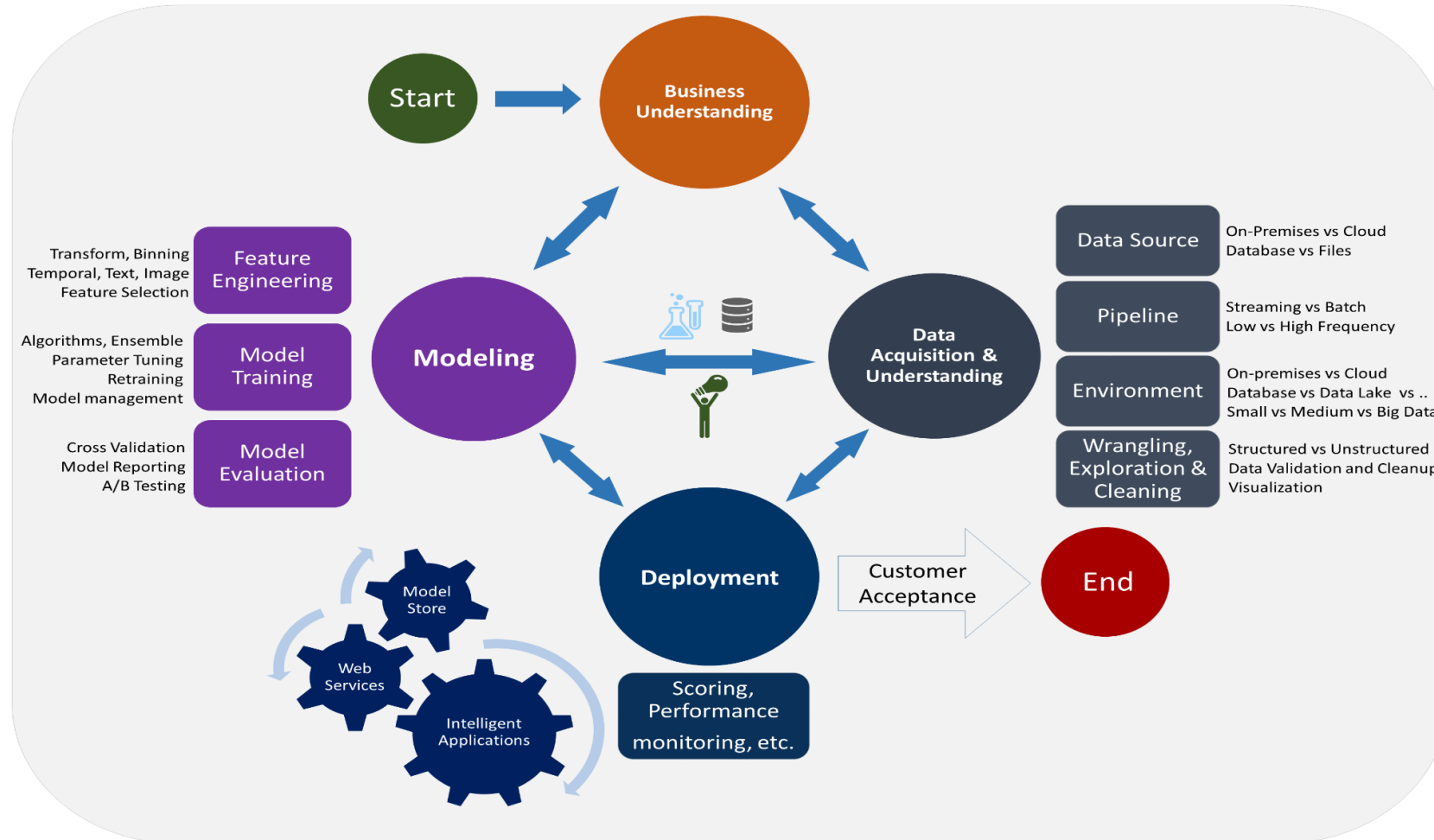
## New Data

Species	Is Independent
Canine	False

## Prediction

Class
Dog

# The Continuous Machine Learning Process



# What is a **model**?

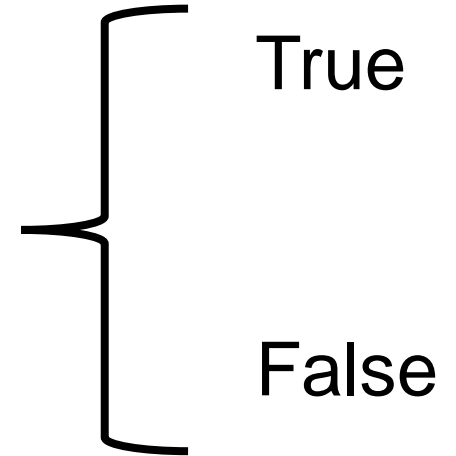


Input



$f(x)$

Model



Output

**Demo**  
**Restaurant Inspections**  
**Training**  
**(ML.NET Automated (Auto) ML)**

**Demo**  
**Restaurant Inspections**  
**Enrichment**  
**(.NET for Apache Spark + SQL)**

# **Demo Restaurant Inspections Enrichment (Azure HDInsight)**

# Takeaways

- Apache Spark is a fast and extensible computing platform for processing Big Data workloads.
- .NET for Apache Spark brings the power of Apache Spark to the .NET ecosystem.
- .NET for Apache Spark provides easy integration with other .NET libraries / frameworks such as ML.NET.



# Resources

- <http://luisquintanilla.me/2019/09/15/restaurant-inspections-etl-data-enrichment-spark-auto-ml-net/>
- <https://docs.microsoft.com/en-us/dotnet/spark/>
- <https://github.com/dotnet/spark>
- <https://docs.microsoft.com/en-us/dotnet/machine-learning/>
- <https://github.com/dotnet/machinelearning>
- <https://spark.apache.org/>

# Questions ?