

# Understanding Databricks/Spark Performance Tuning

## Lesson 01: The Spark Architecture & Bottlenecks

by Bryan Cafferky from my YouTube channel



# Where We're Going?

- **Databricks/Apache Spark Architecture**
- **Performance Bottlenecks**
- **Tuning Options Vary By Platform**
- **Wrap Up**



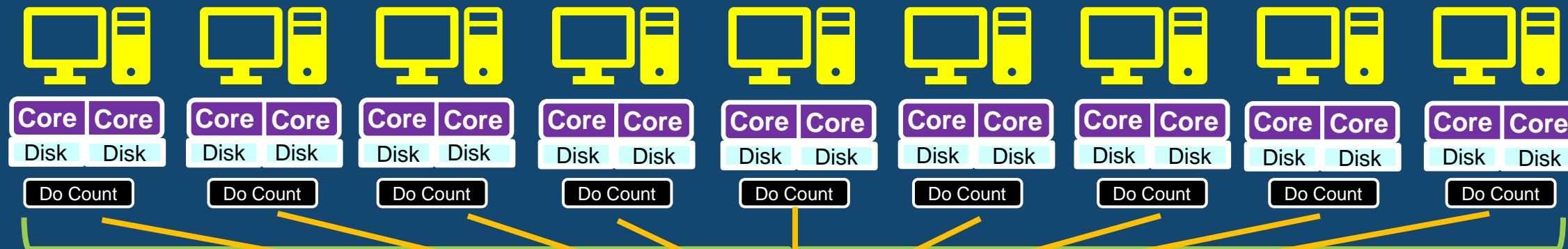
databricks

# Spark Cluster



External Storage

Worker Nodes



Send Back Result

Partition Data by City  
and Copy One City to Each Node Executor

Distributing the Data Over the Cluster

## Constraints

- Hardware/Resources
- Software (Spark/DBR)
- Environment Configurations
- Your Code/Application
- Data Source & Format
- Data Distribution



```
SELECT City, Count(*) FROM PhoneBook  
Group By City  
Order By City
```



Phone Book

# Options Vary By Platform

**Delta Lake vs. Non Delta Lake**

**Open-Source Spark**

**Databricks**

# Wrapping Up

- Databricks/Apache Spark Architecture
- Performance
- Tuning Options vary By Platform

Thank You!