Apache Spark Workshop

WIT NYC Hour of Code | 7 Dec 2017

Welcome!

- Today's Agenda
  - What is distributed computing?
  - What is Spark and why does it matter?
  - What is Databricks?
  - Hands-on Tutorial using Spark on Databricks

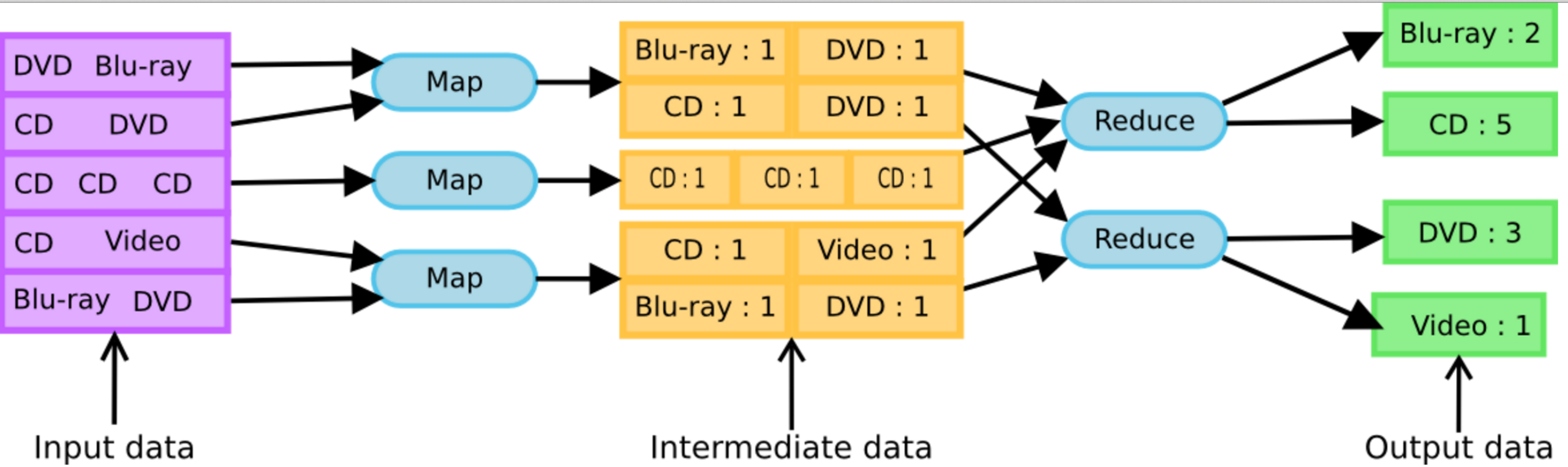
What is distributed computing?

- Computing across clusters
- Scalable, fault-tolerant
- Foundation of Hadoop

What is Spark and why is it important?

- Cluster computing framework for big data processing
- Spark APIs to code in Python (PySpark), R, Java, Scala
- SparkSQL for relational data querying
- MLlib for machine learning
- GraphX for graph processing

What is MapReduce?



Spark versus Hadoop MapReduce

- Spark can be 100x faster than MR due to in-memory procesing (contrast with MR storing to disk)
- Map-reduce concepts still exist in Spark!

Counting in MapReduce

```
def mapper(line):
    words = line.split()
    for word in words:
        yield word, 1

def reducer(word, counts):
    print word, sum(counts)
```

Counting in Spark

```
wordCounts = textDocument \
    .flatMap(lambda line: line.split()) \
    .map(lambda word: (word, 1)) \
    .reduceByKey(lambda x, y: x + y)
```

Some deeper Spark concepts…

- **Spark 1.x**
  - **Spark Context** (SC): must be created at the start of Spark session
  - **Resilient Distributed Dataset** (RDD): data across cluster nodes that can be acted on in parallel
    - new RDDs are created lazily with each transformation, such as *map*, *reduceByKey*, etc
    - can be converted to/from Spark's relational **DataFrames**
  - **SQL Context**: created from SC and provides RDMS operations
- **Spark 2.x**
  - **SparkSession**: A unified entry point for manipulating data with Spark

Some deeper Spark concepts (con't)

- **Transformation**
  - Operations that will not be completed at the time you write and execute the code in a cell
  - They will only get executed once you have called a **action**
  - Example: select, sum, filter
- **Action**
  - Operations that are computed by Spark right at the time of their execution
  - They consist of running all of the previous transformations in order to get back an actual result
  - Example: show, count, save

What is Databricks

- Databricks is a managed platform for running Apache Spark
  - No need to manage complex cluster
- Spark was originally written by the founders of Databricks during their time at UC Berkeley
- The community version provides the Workspace with a free mini 6GB cluster

Hands-on Tutorial: ”Hello, Spark!”

- Take out your laptops (and/or share with a neighbor!)
- Head to: https://community.cloud.databricks.com/
  - Register an account
  - Import the notebook
    - The Lab: http://bit.ly/2BaX4f0
    - The Solutions: http://bit.ly/2B5NI44
    - Click on “Import Notebook” on the top right corner
- Questions? Let us know!