databricks

PySpark

1. DATABRICKS & PYSPARK

INTRODUCTION

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

Databricks features

- Databricks platform provides a number of features including Clusters, Data engineering, Machine Learning, Lakehouse, Generative AI and other.

CDU
Cloud And Data Universe

Databricks features

- Databricks is the go to choice for Data engineering which includes PySpark development.
- Databricks provides notebooks for Spark & SQL Development.

Cloud And Data Universe

What you will learn in this section?

- Databricks: Generic features like DBFS, dbutils, mount ADLS, calling notebooks, widgets, jobs, etc.
- PySpark: In detail from start to end.

Pre-requisites

- SQL
- Python Fundamentals
- Databricks community Edition

CDU
Cloud And Data Universe

databricks

PySpark

2. DATABRICKS & PYSPARK

INTRODUCTION TO APACHE SPARK

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

Apache Spark

- Apache Spark is a processing engine.
- It is widely used in Data engineering, Data science and Machine learning.
- Apache spark is an alternative to Map reduce.

CDU
Cloud And Data Universe

Mapreduce Challenges

- MapReduce programs are mainly written in Java.
- Development in MapReduce was lengthy.
- Intermediate computed results need to be written to disk for processing of next stage, this was a major drawback as it consumed more time.

CDU
Cloud And Data Universe

How did it begin?

- Experts who worked on Hadoop understood the issue with MapReduce due its limitations.

- They started working on a project to overcome MapReduce limitations and named the project as Spark.

Spark Journey

- Initial development on Spark began around 2009.
- Spark was a game changer in Big data computing!
- The biggest advantage over MapReduce was Spark stored intermediate results in-memory, thereby providing 10 to 100 faster execution as compared to MapReduce!
- First version of Apache Spark was released in 2014.

CDU
Cloud And Data Universe

Databricks

- Creators of Apache Spark later formed a company named Databricks in 2013 and a service offering with same name.
- Databricks service offers Data engineering, Machine learning and Lakehouse solutions.
- Free usage on Databricks community edition.
- Available on cloud: Azure, AWS & GCP.

CDU
Cloud And Data Universe

What does Spark deliver?

- Spark is a processing engine / technology used to perform ETL.
- Spark development involves complete coding.

Cloud And Data Universe

databricks

PySpark

3. DATABRICKS & PYSPARK

SPARK COMPONENTS & API

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

Spark Components

Spark SQL

Spark MLlib

Spark Structured Streaming

GraphX

CDU
Cloud And Data Universe

Spark API's

Scala

Python

Java

R

SQL

CDU
Cloud And Data Universe

databricks

PySpark

4. DATABRICKS & PYSPARK

SPARK ARCHITECTURE

Follow us on :

YouTube

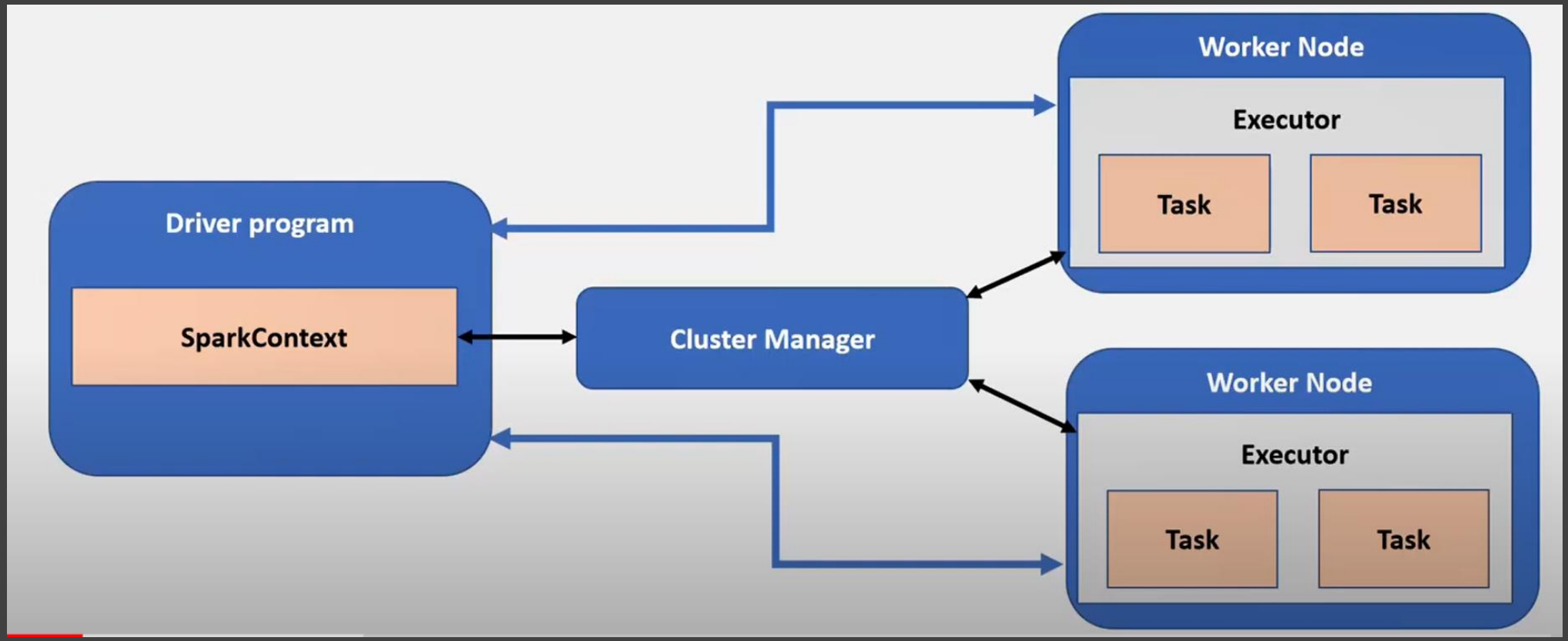
LinkedIn

CDU

Cloud And Data Universe

databricks PySpark

Spark Architecture



CDU
Cloud And Data Universe

Spark Architecture

- Spark uses master-slave architecture.
- The main task of spark is to distribute data across cluster and process it in parallel over nodes.

Spark Application

- Spark Application is a program written by user.
- It consists of driver program and executors.

Driver Program

- Driver program initiates the execution of program.
- Its runs the main() function of application.
- It creates the SparkContext.

SparkContext & SparkSession

- SparkContext is an entry point to spark.
- By using SparkContext we can create a RDD which is fundamental unit of storage in spark.
- After initial version of spark, SparkSession was introduced which became entry point to spark.
- SparkSession includes SparkContext, SQLContext, HiveContext and StreamingContext.

CDU
Cloud And Data Universe

Cluster Manager

- Cluster Manager is responsible for acquiring resources in a cluster.
- The driver program requests for resources to the cluster manager.
- Then cluster manager launches executors on worker nodes as requested by driver program.
- Cluster managers: standalone, Mesos, Yarn

Execution Modes

- Cluster Mode: Driver is launched inside the cluster.
- Client Mode: Driver is launched outside the cluster i.e. on client machine from which spark application was submitted.
- Local Mode: Application runs on single machine.

CDU
Cloud And Data Universe

Executors

- Executor is a java process launched on worker node.
- Executors register themselves with driver program at the beginning.
- The executors are dynamically added or removed during the task execution.

Task

- Task is a unit or chunk of data sent to executor.
- Each executor runs one to many tasks

Job

- Job is a process of parallel computation.
- It involves computation of multiple tasks.

Cloud And Data Universe

databricks

PySpark

5. DATABRICKS & PYSPARK

RDD

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

RDD

- RDD stands for Resilient distributed dataset.
- It is the fundamental unit of storage in spark.
- We can create a RDD using SparkContext.
- RDD is immutable.
- RDD is partitioned across worker nodes.

CDU
Cloud And Data Universe

RDD

- RDD stands for Resilient distributed dataset.
- Resilient: Relates to fault-tolerance i.e. ability to recover from failure.
- Distributed: Partitioned across nodes.
- Dataset: Collection of records which is stored in files like csv, json, etc.

CDU
Cloud And Data Universe

RDD Operations

- Once RDD is created you can perform different operations on it.
- There are broadly 2 types of operations.
- Transformations & Actions.

RDD Transformations

- Once we can create and RDD, we can perform various transformations as needed.
- These can be mostly done using map(), filter(), reducebykey(),etc.
- Remember, each time we apply a transformation on a RDD, we will get a new RDD, existing RDD will remain unchanged as RDD is immutable.

Shuffle

- Data is distributed across nodes initially.
- Once the data is read we need to perform some operations on it.
- Certain operations require data to be re-distributed across the nodes.
- This triggers an event called as shuffle.
- Shuffle involves copying of data across the nodes.
- Shuffle is a costly operation!

CDU
Cloud And Data Universe

Types of Transformation

- 2 types of transformations: Narrow & Wide.
- In Narrow transformation shuffle doesn't occur as there is no need to re-copy the data across worker nodes in intermediate steps.
- In wide transformation shuffle occurs as there is need to re-copy the data across worker nodes in intermediate steps.

Actions

- Once all the transformations are done we need to call an action for the result to be computed.

- It is important to note here, spark doesn't use any resources or initiate any computation unless an action is called!

Scenario

1. Suppose we want to read data from a text file.

2. Next, we need to perform few operations on data like filter and aggregate.

3. Fetch final results.

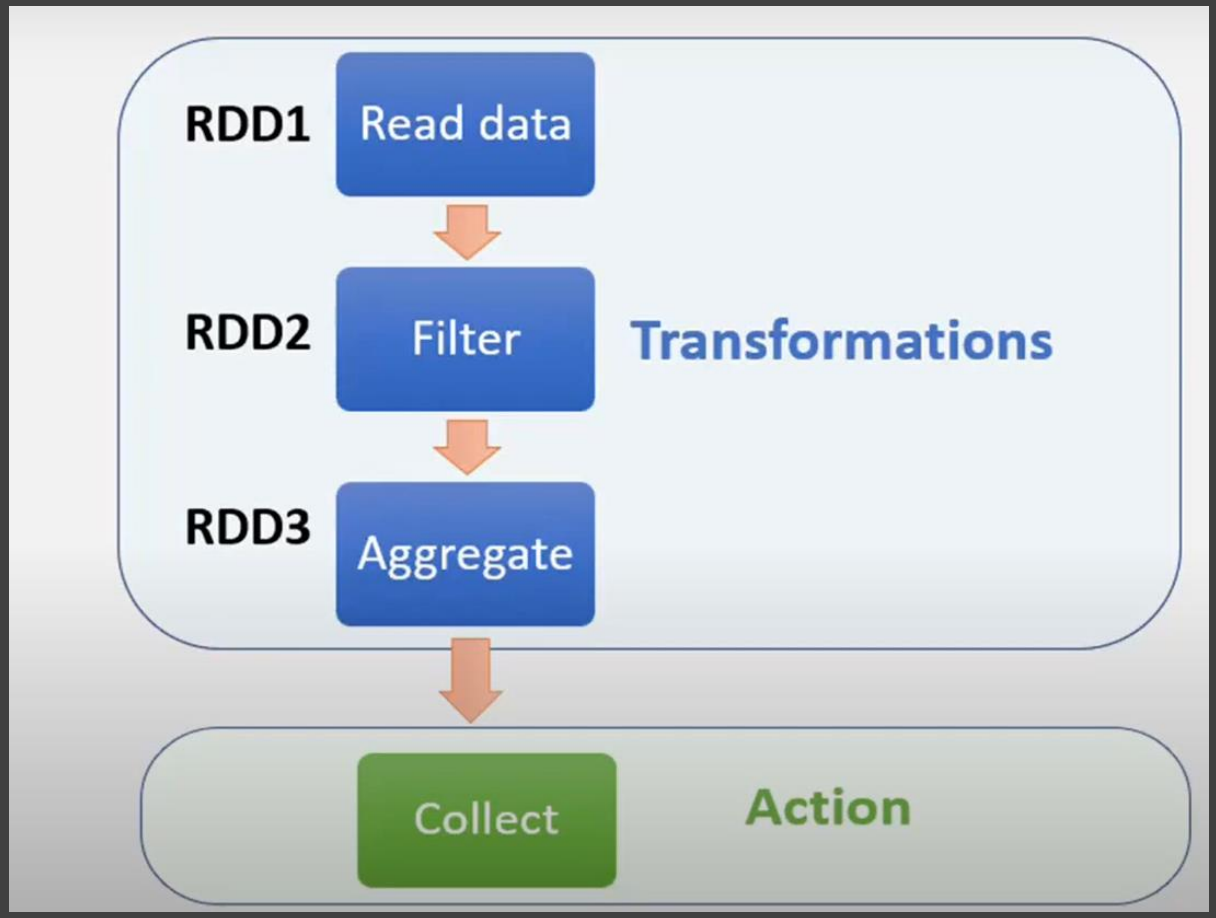
Important points

- Operations like filter, aggregate, etc. are transformations.
- When spark runs the transformations, actual computation doesn't take place, instead it records these entries in lineage graph.
- Lineage graph contains flow of rdd's pointing to its parent rdd's.

CDU
Cloud And Data Universe

databricks PySpark

Lineage graph



CDU
Cloud And Data Universe

databricks

PySpark

6. DATABRICKS & PYSPARK

CREATE RDD FROM LIST

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

7. DATABRICKS & PYSPARK

CONTROL PARTITIONS IN RDD

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

8. DATABRICKS & PYSPARK

CREATE RDD FROM TEXTFILE

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

9. DATABRICKS & PYSPARK

FLATMAP, MAP , REDUCEBYKEY TRANSFORMATIONS ON RDD

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

10.DATABRICKS & PYSPARK

LINEAGE GRAPH

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

11. DATABRICKS & PYSPARK

UNDERSTANDING DAG FUNDAMENTALS

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

12. DATABRICKS & PYSPARK

MAPREDUCE WORKING

Follow us on :

YouTube

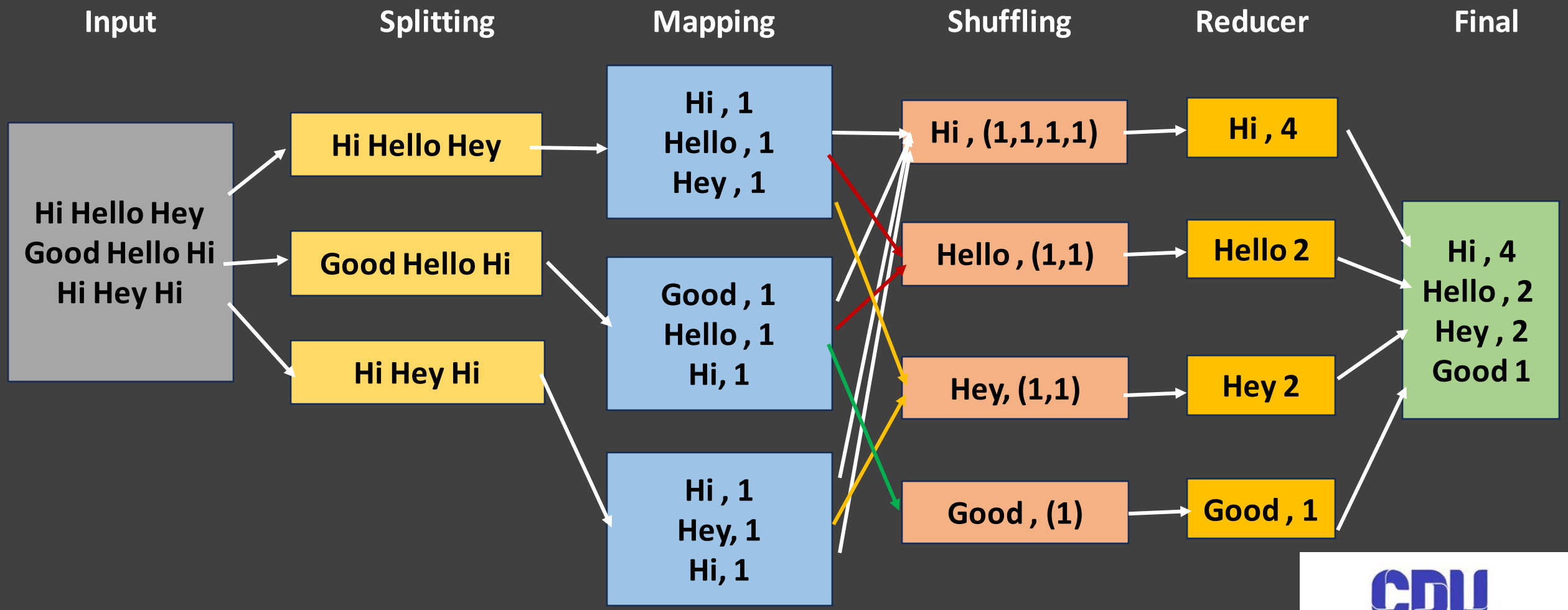
LinkedIn

CDU

Cloud And Data Universe

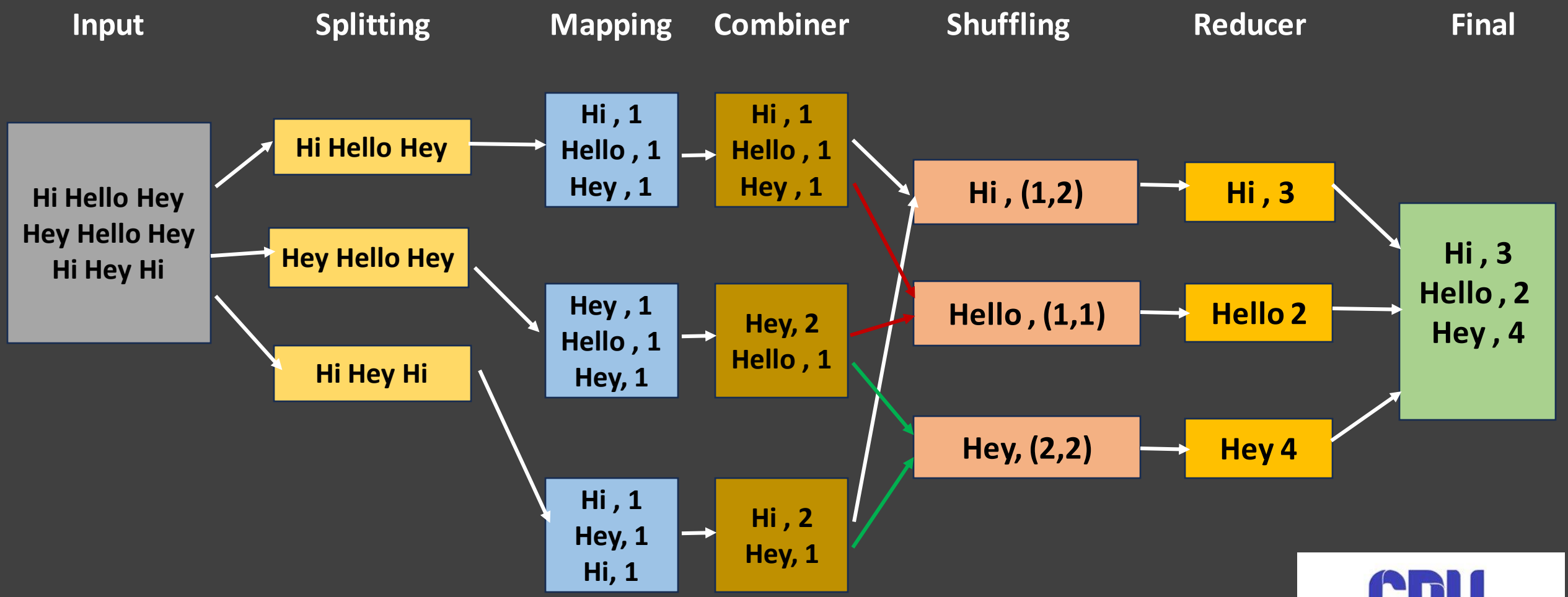
databricks PySpark

Map Reduce – Wordcount example



CDU
Cloud And Data Universe

Map Reduce – Wordcount example with Combiner



Cloud And Data Universe

databricks

PySpark

13. DATABRICKS & PYSPARK

REDUCEBYKEY VS REDUCEBYKEYLOCALLY

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

14.DATABRICKS & PYSPARK

GROUPBYKEY

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

15. DATABRICKS & PYSPARK

FILTER TRANSFORMATION ON RDD

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

16. DATABRICKS & PYSPARK

SORTBY & SORTBYKEY TRANSFORMATIONS ON RDD

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

17. DATABRICKS & PYSPARK

EXTRACT TOP BOTTOM FROM RDD

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

databricks

PySpark

18. DATABRICKS & PYSPARK

SAVE RDD AS TEXTFILE

Follow us on :

YouTube

LinkedIn

CDU

Cloud And Data Universe

Brought to you by:

Follow us on :