Big Data Modeling & Analytics

(course outline)

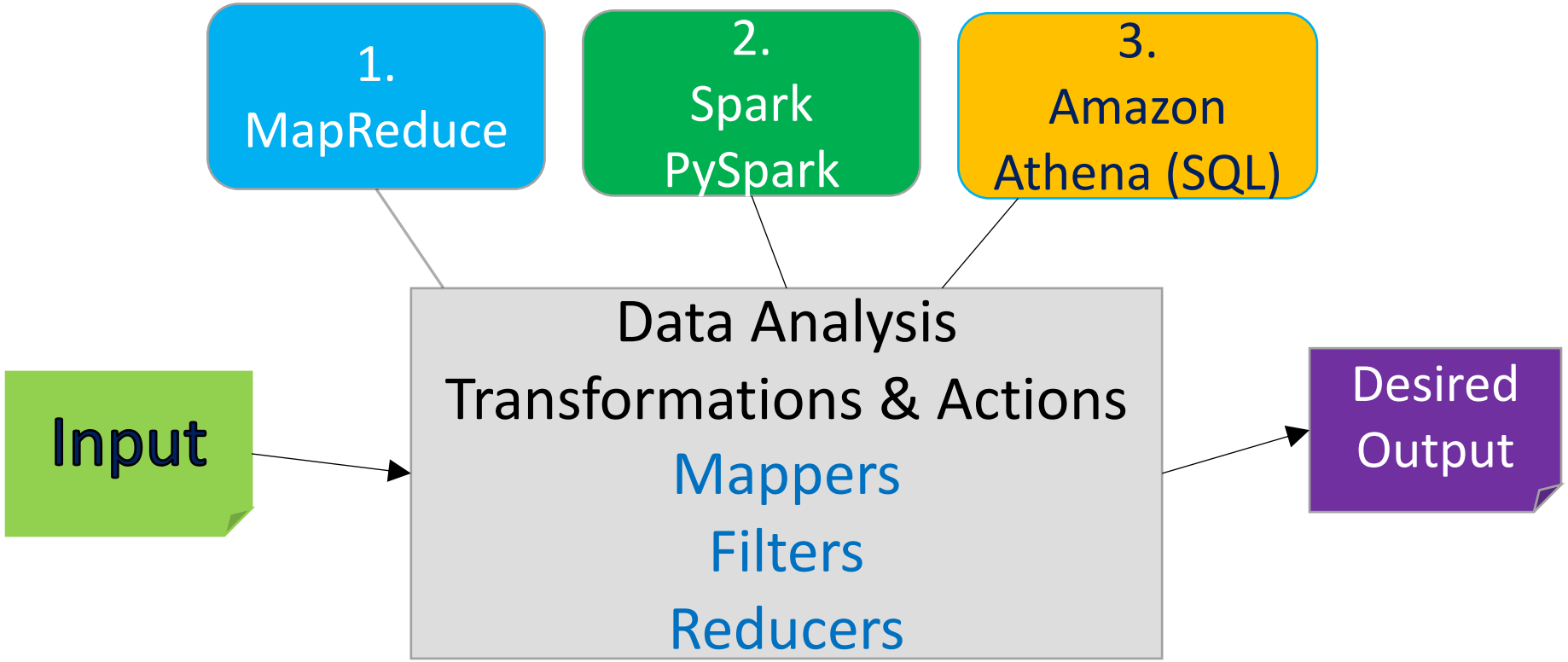
Mahmoud (Max) Parsian

Ph.D. in Computer Science

Main Course Components

1. Introduction to MapReduce Paradigm (25%)
2. Spark and PySpark (60%)
3. Serverless SQL Access to Big Data (15%)
    - Amazon Athena
    - Google BigQuery
    - Snowflake

Data Analysis



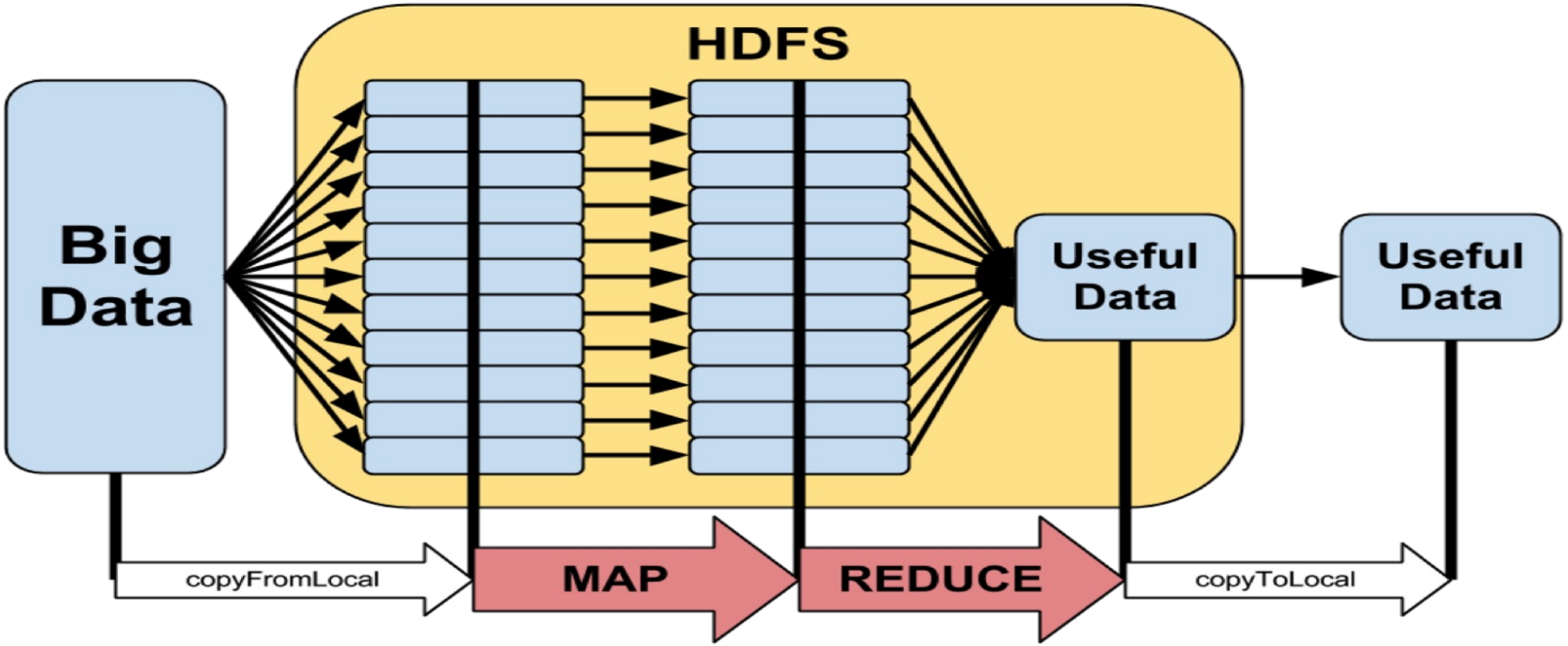
MapReduce Paradigm

- **MapReduce** is a programming **paradigm/model** that enables **massive scalability** across hundreds or thousands of servers in a cluster.
- MapReduce has 3 functions:
  - `map()`
  - `reduce()`
  - `combine()`

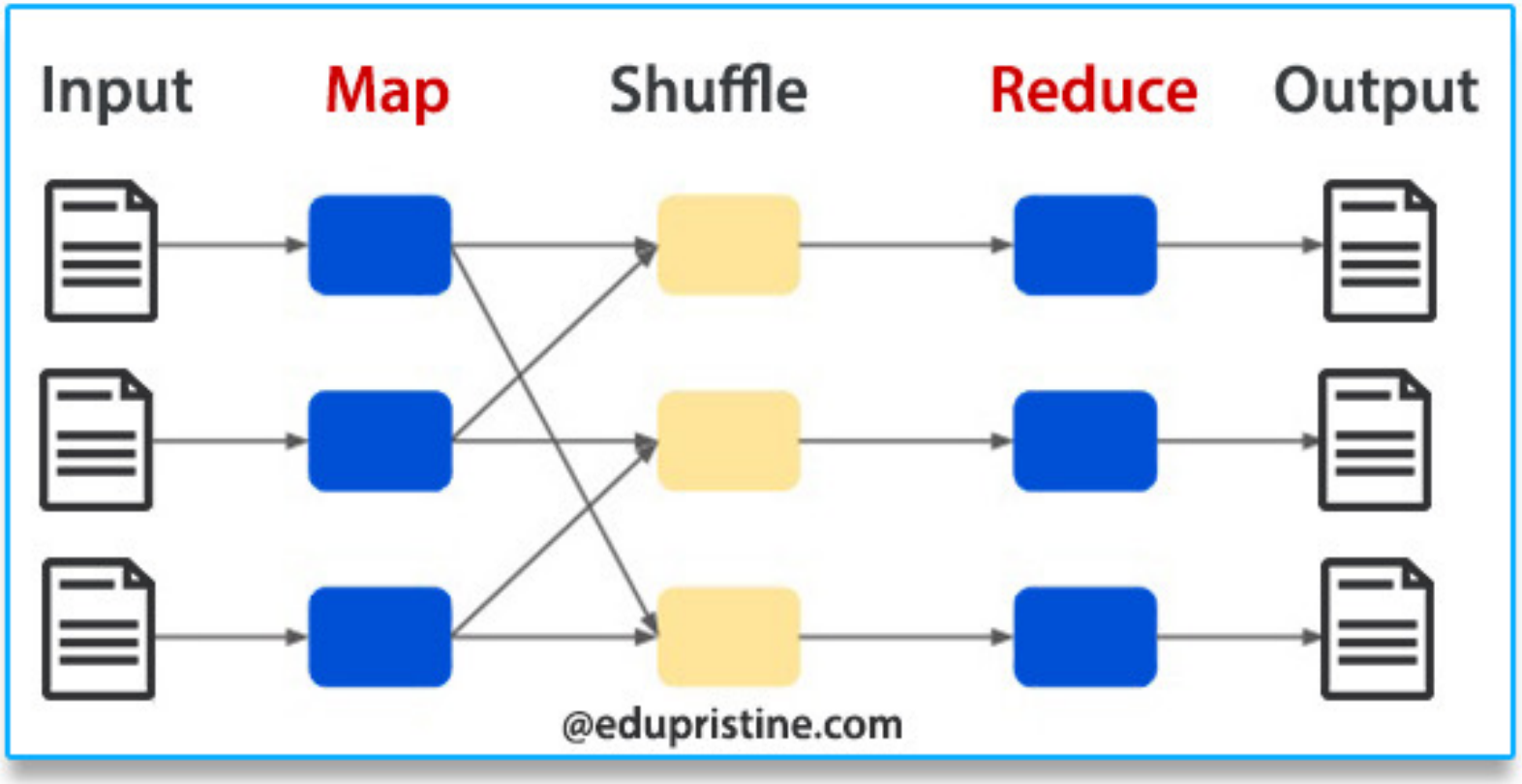
MapReduce Paradigm

- **MapReduce is a foundation model/paradigm for distributed computing using clusters**
- **MapReduce implementations:**
  - Apache Hadoop implements MapReduce
  - Apache Spark implements superset of MapReduce
  - Apache Tez implements MapReduce

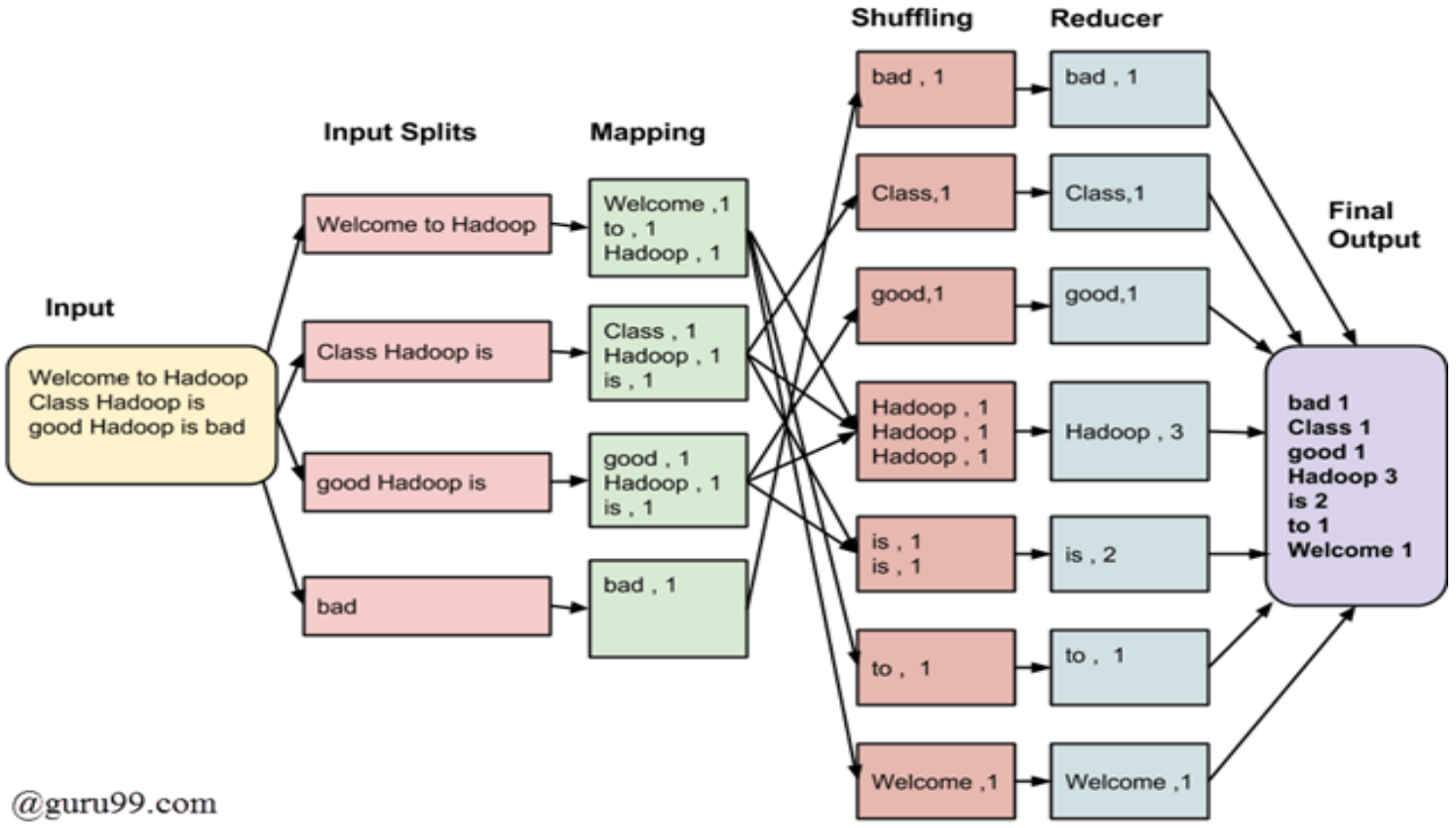
MapReduce Paradigm



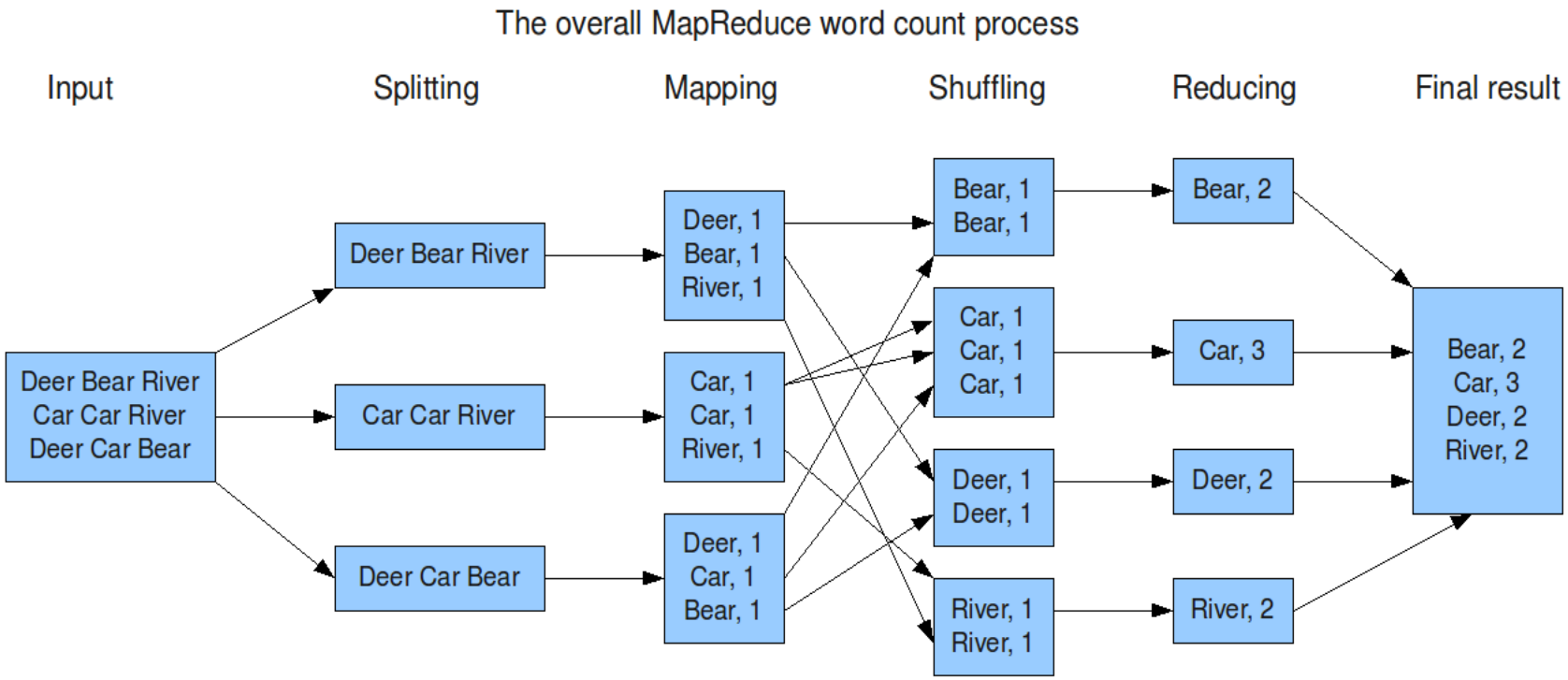
MapReduce Paradigm



MapReduce Paradigm Example-1



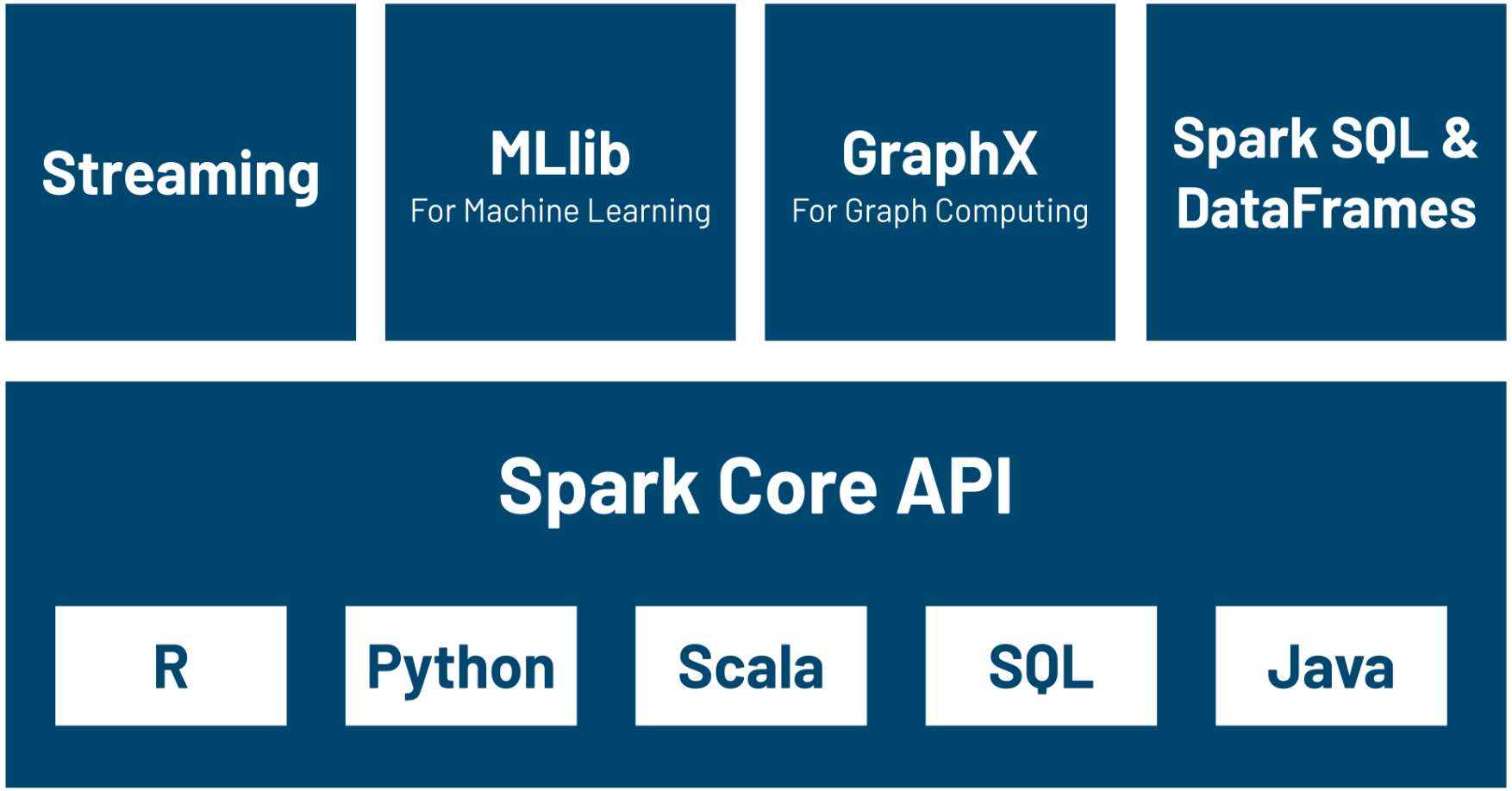
MapReduce Paradigm Example-2



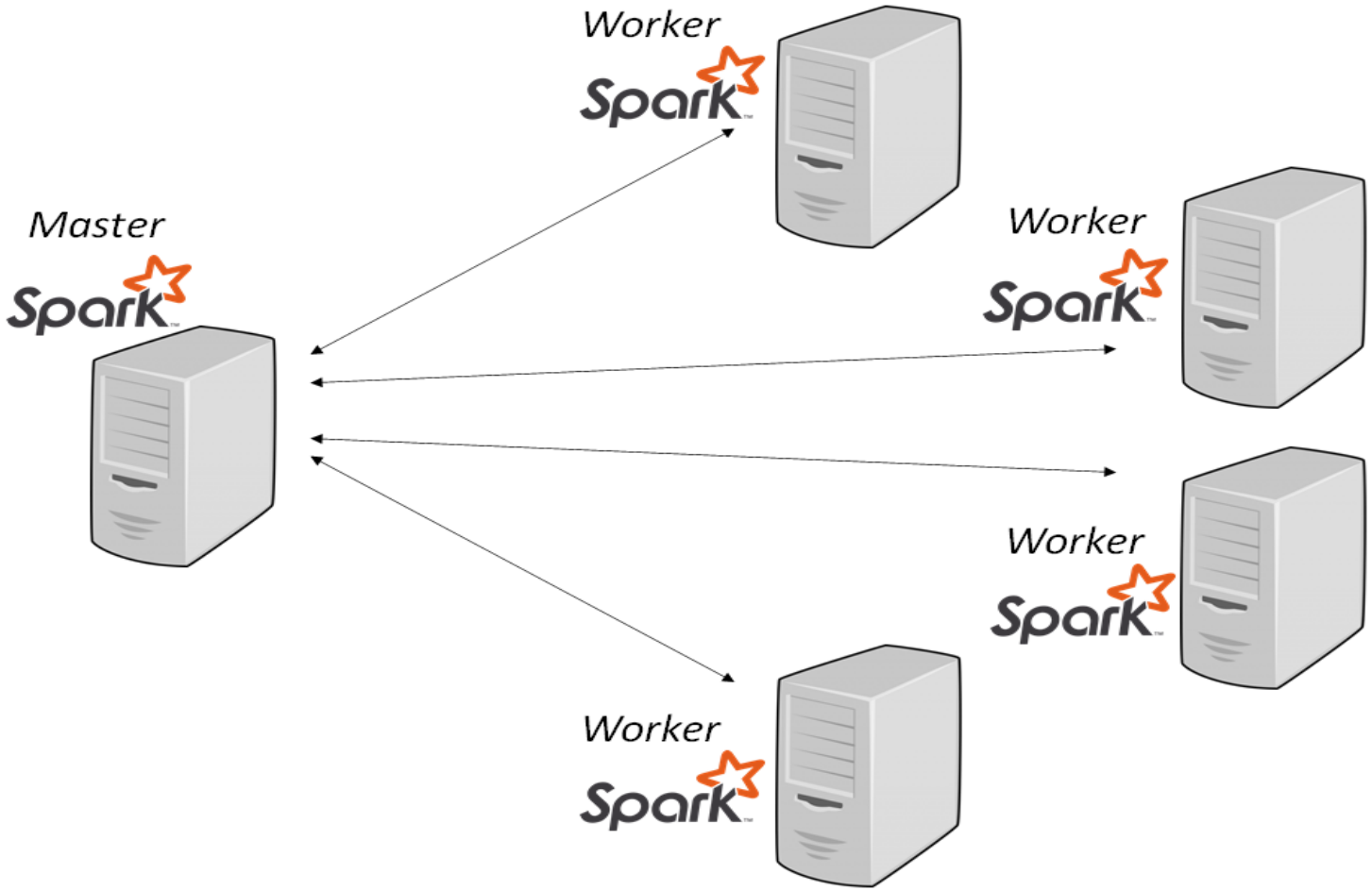
Apache Spark

- Apache **Spark** is a multi-language (Java, Python, Scala) engine for executing data engineering, data science, and machine learning on single-node machines or clusters.
- PySpark is a Python API for Spark
- Spark is a superset of MapReduce

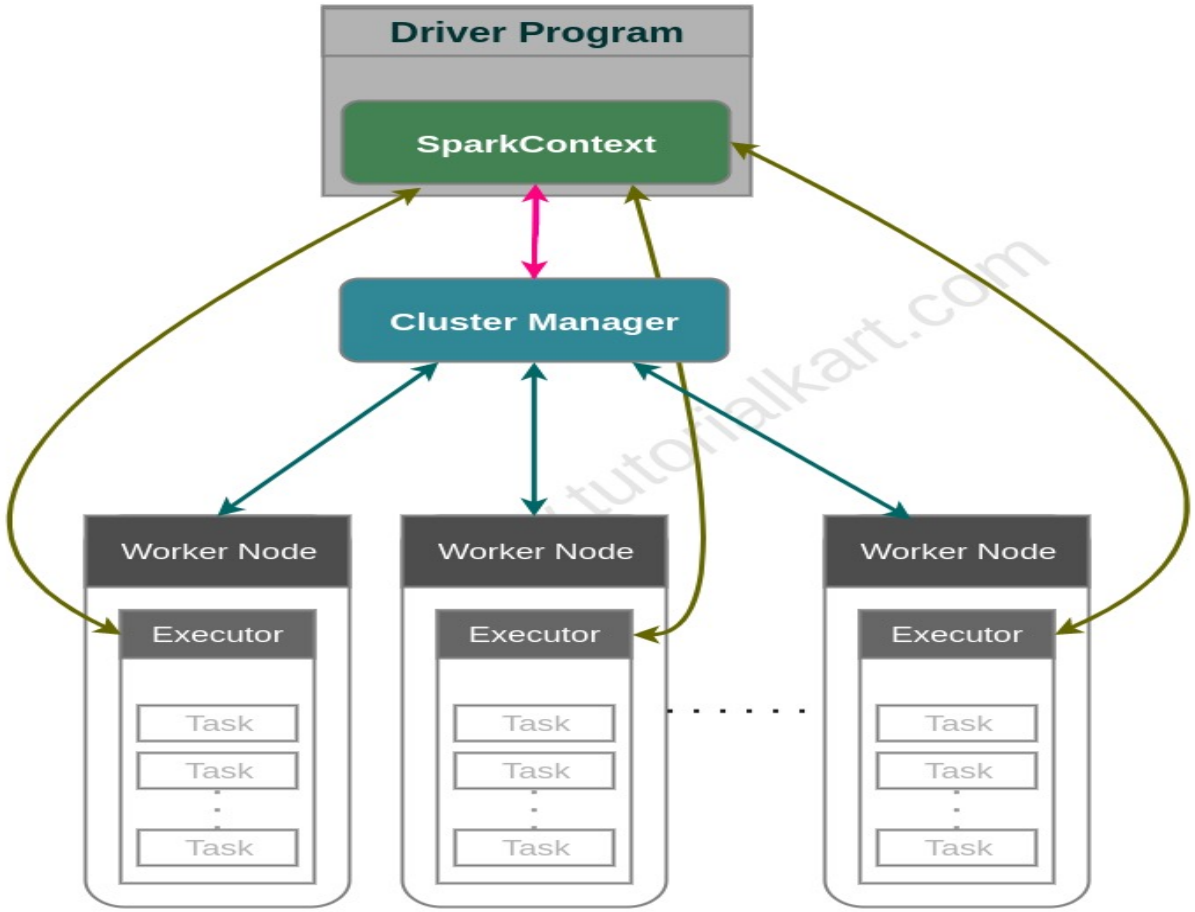
Apache Spark: Components



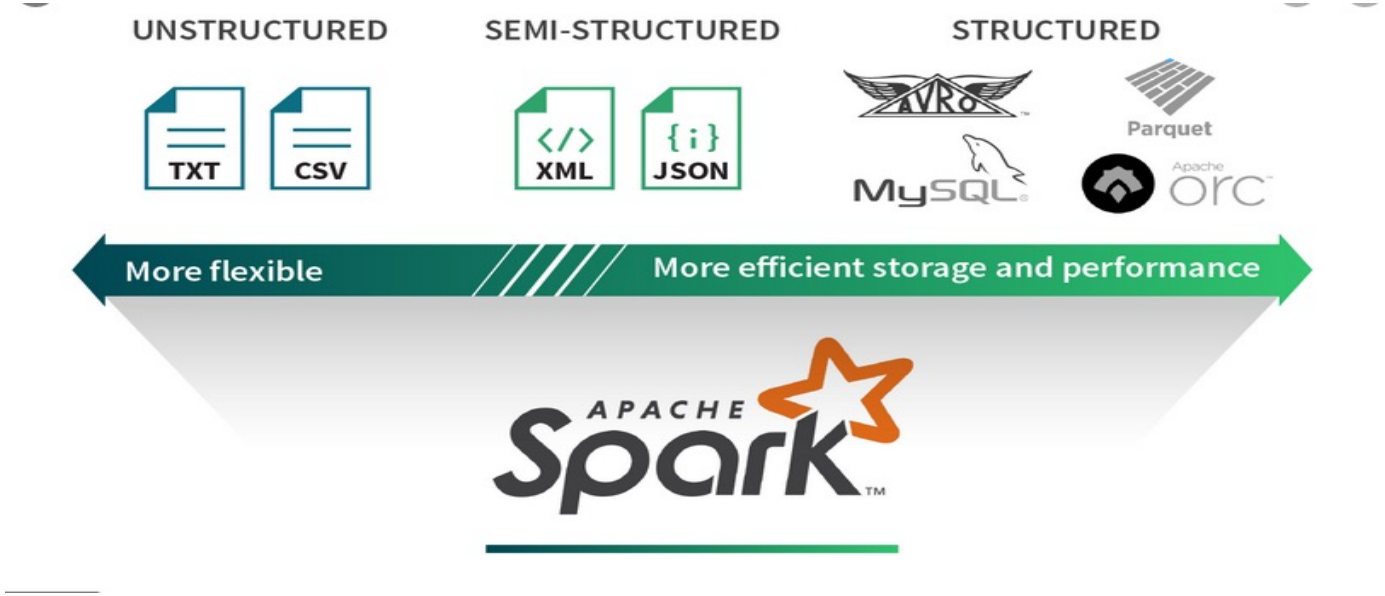
Apache Spark: runs in a cluster



Apache Spark: Cluster Manager



Spark Data Abstractions: Read/Write from/to Many DataSources



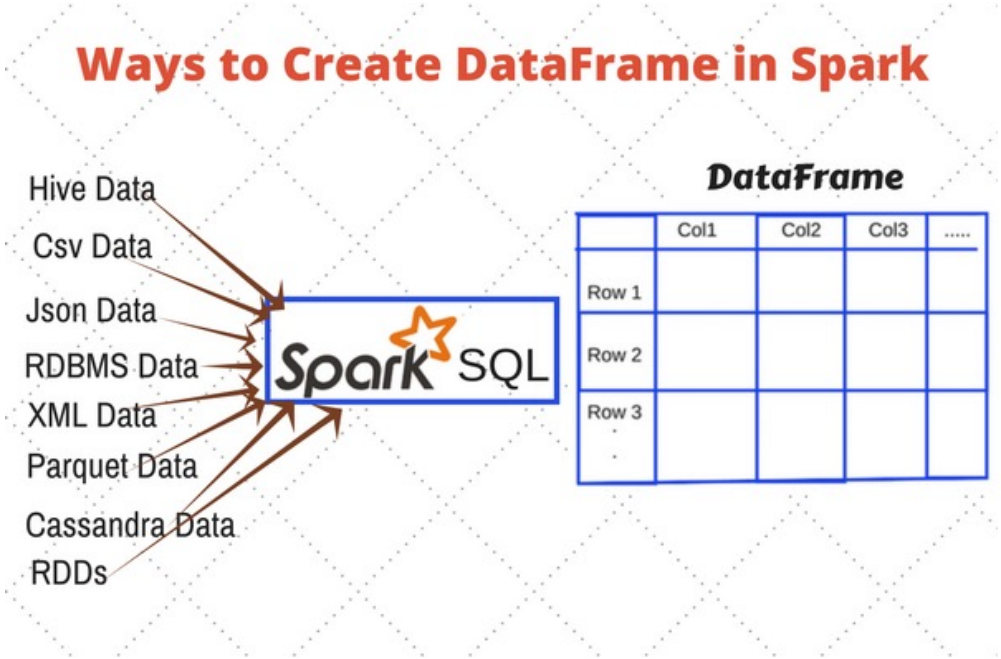
Spark Data Abstractions

- **Resilient Distributed Datasets (RDD)**
  - Billions of data points (elements/records)
- **DataFrame**
  - Data is represented as a table of rows and named columns
  - Billions of rows of data

Spark Data Abstractions: RDDs
Billions of elements

Element-1:	`(gene-1, (3.0, 4.5))`
Element-2:	`(gene-2, (1.0, 1.5))`
Element-3:	`(gene-1, (2.1, 1.6))`
	…
	`(gene-8, (3.1, 5.1))`

Spark Data Abstractions: DataFrame



Amazon Athena

- Interactive query service
- Serverless. Zero infrastructure. Zero administration.
- Put your data in S3
- Access your data by SQL
- Pay by query
- Fast performance

Amazon Athena