



Introduction to Big Data Technologies

Anurag Nagar, Ph.D.

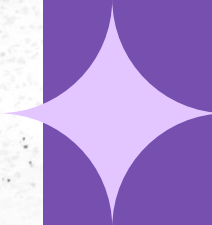




Table of contents

01

**What is
Big Data**

02

**Cluster
Computing**

03

**Spark
Coding**

04

Demo






01



What is Big Data



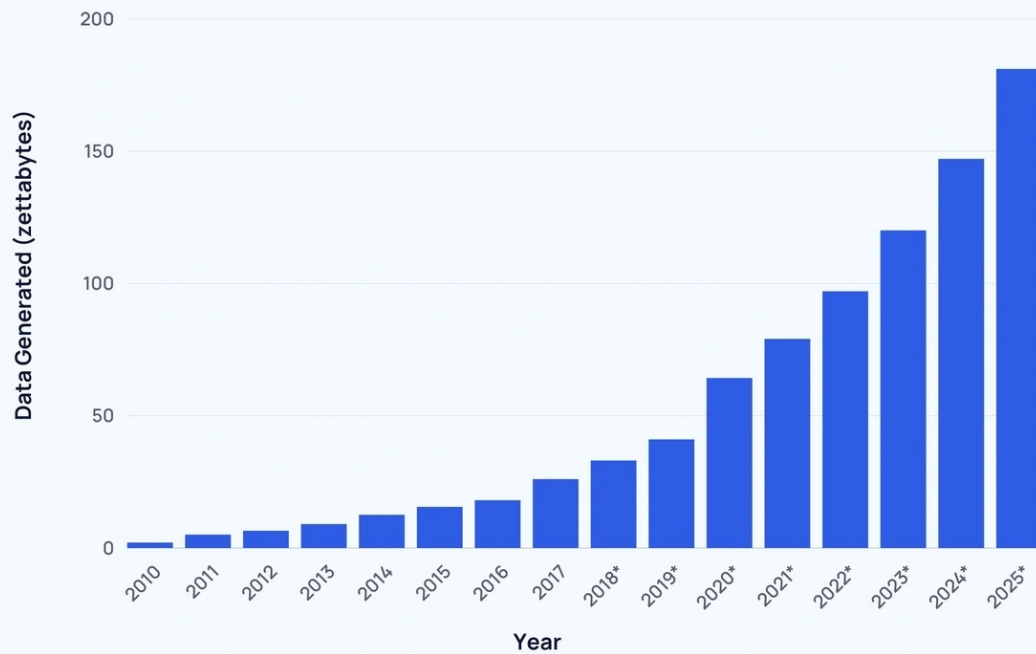
Big Data

- Data is all around you.
 - In recent years there has been a shift in the type of data:
 - Structured -> Unstructured
 - Fixed, pre-determined units -> Variable units
 - Smaller size -> Very large sizes
 - Lot of time for analysis -> Instant analysis
- 



How Big is Big Data?

Global Data Generated Annually



Source: Statista





Who Produces Big Data

- Social Networks
- Media
- Telecom Companies
- Healthcare and Medicine
- Large science projects
- Each one of us!





Defining Big Data

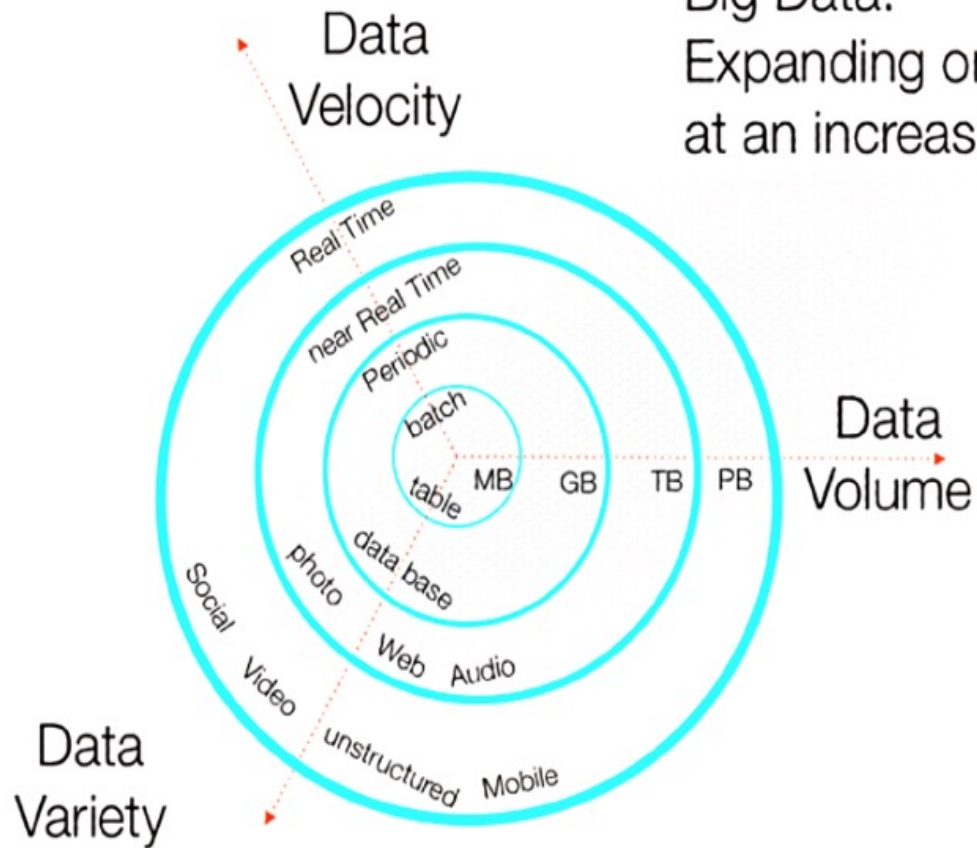
Data growth challenges and opportunities are three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources).

- **Doug Laney of Gartner group**



Defining Big Data

Big Data:
Expanding on 3 fronts
at an increasing rate.





V's of Big Data

Characteristics of Big Data.

- Volume
- Velocity
- Variety
- Veracity
- Variability
- Value

More details at:
<https://cloud.google.com/learn/what-is-big-data>





Big Data

- Big Data produced by major entities is in raw form.
- Need to extract value from this raw data.
- This is where the field of analytics, and data mining come into play.
- Need a way to store and process this data inexpensively.
- Cluster computing solves this issue



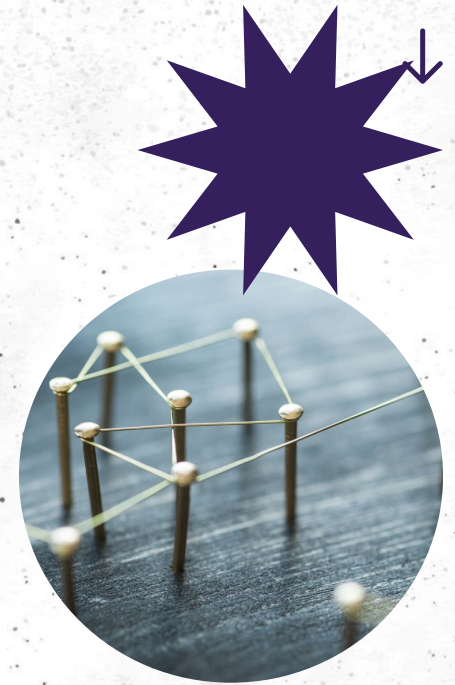


02

Cluster Computing

Cluster Computing

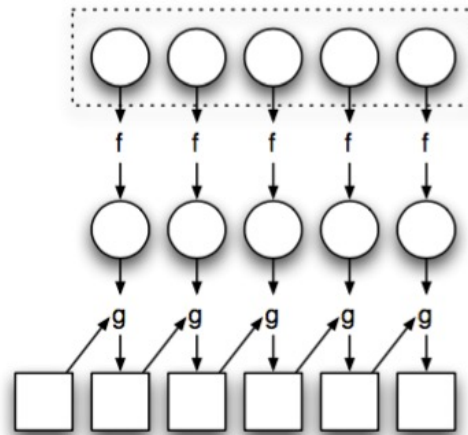
- Set of computers (nodes) connected together and working in sync.
- Distributed Computing
- Different than multi-core computing
- Higher availability
- Higher processing power
- Cheaper as compared to multi-core computing
- The newest manifestation of cluster computing is cloud computing.





How to Program a Cluster

- MapReduce programming
 - **Key feature: higher order functions**
 - ▶ Functions that accept other functions as arguments
 - ▶ **Map** and **Fold (Reduce)**



f is applied to every element and it results in a new list

g starts with an initial value and reduces every element i.e. compacts list to a scalar

Figure: Illustration of *map* and *fold*.



Map Operation

- Define a function: `square x = x * x`

- Apply on a list: `>>> map square [1, 2, 3, 4, 5]`

- Get another list: `[1, 4, 9, 16, 25]`,

Reduce (Fold) Operation

- Define an operator: +
- Initial value = 0
- Apply on a list: `[1, 2, 3, 4, 5]`
- Get a scalar: 15

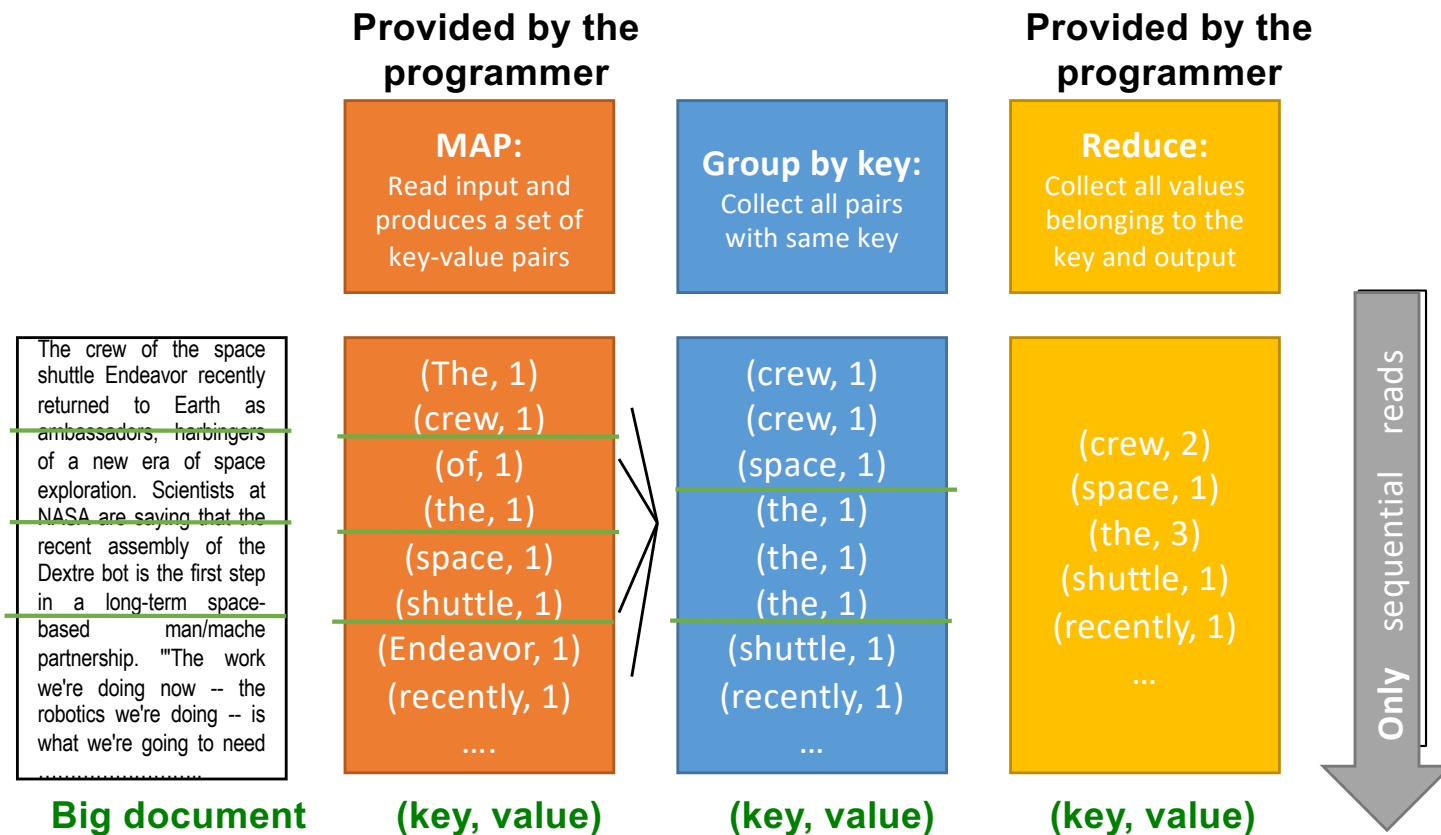
Example – Word Count

Programming Model: MapReduce

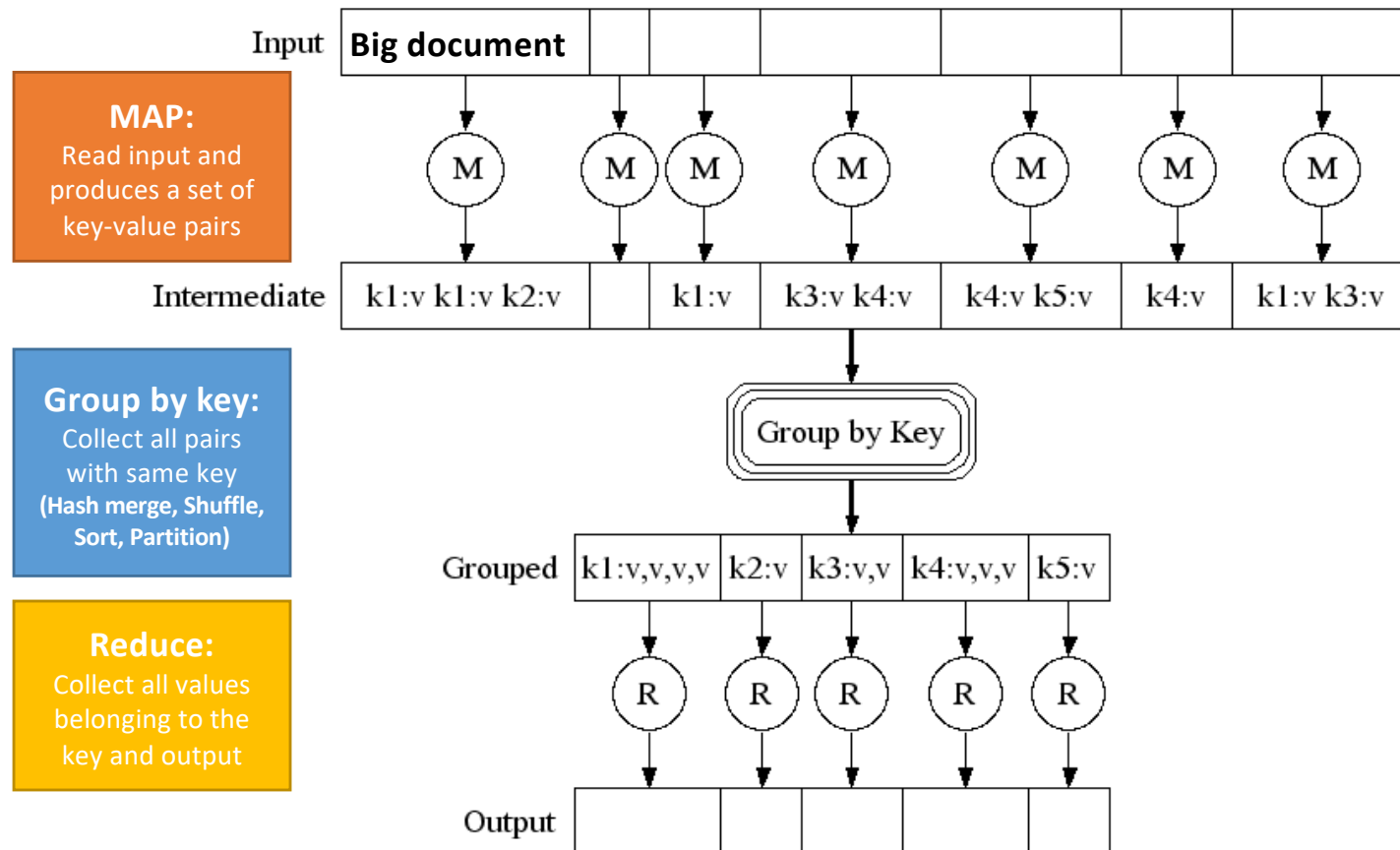
Warm-up task:

- We have a huge text document
- Count the number of times each distinct word appears in the file
- **Sample application:**
 - Analyze web server logs to find popular URLs

MapReduce: Word Counting



Map-Reduce: A diagram



Word Count Using MapReduce

map(key, value):

```
// key: document name; value: text of the document
for each word w in value:
    emit(w, 1)
```

reduce(key, values):

```
// key: a word; value: an iterator over counts
result = 0
for each count v in values:
    result += v
emit(key, result)
```


Map-Reduce: Environment

Map-Reduce environment takes care of:

- **Partitioning** the input data (input splits)
- **Scheduling** the program's execution across a set of machines
- Performing the **group by key** step
- Handling machine **failures**
- Managing required inter-machine **communication**



03

Spark Coding

Spark Coding

- We will use Apache Spark as the coding environment.
- Steps:
We will use Google Colab for coding in Spark:
colab.research.google.com
- PySpark notebook can be downloaded from
github.com/a-nagar/big_data





Thanks!

Do you have any questions?

anurag.nagar@utdallas.edu

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

