HADOOP INTRODUCTION

OUTLINE

- Homework Discussion
- Course Project (Option-1) Discussion
- Hadoop Framework
  - Motivations
  - Challenges
  - Hadoop vs. Grid Computing
  - Hadoop Main Components
  - Hadoop Software Architecture

HOMEWORK DISCUSSION

MOTIVATIONS

- Big Data!
- Storage and Analysis
  - Storage capacity increases faster than the access speeds

Year	HD Size	Transfer Speed (MB/s)	Time to read the whole drive
1990	1,370 MB (1GB)	4.4 MB/s	~ 5 minutes
2000	1 TB or 1000 GB	100 MB/s	~ 2.5 hours

  - Need parallel data access to get thing done quickly
    - 1 machine is accessing 1000 GB is much slower than 100 machines, each is accessing 10 GB.
  - Shared access for efficiency and scalability

CHALLENGES

- Analysis tasks need to combined data from multiple sources
  - Need a paradigm that transparently split and merge data
- Challenge of parallel data access to and from multiple disks
  - Hardware failure

WHY HADOOP?

- **Open-source** software framework for storing data and running applications on clusters of **commodity hardware**.
  - Cheap to implement and expand
- Provide **massive storage** for any kind of data, enormous processing power and the ability to **handle virtually limitless concurrent tasks or jobs**
  - Scalable

HADOOP VS GRID COMPUTING

- Existing Grid Computing, ie. HPC
  - Distribute tasks to process data in a shared file system
  - Data need to move to the machines that run tasks
  - Not suitable for tasks accessing large data volumes

- Hadoop
  - Try to **co-locate** the data with the computing node
    - → Data Locality
  - Avoid copying data around
  - Automate fault recovery

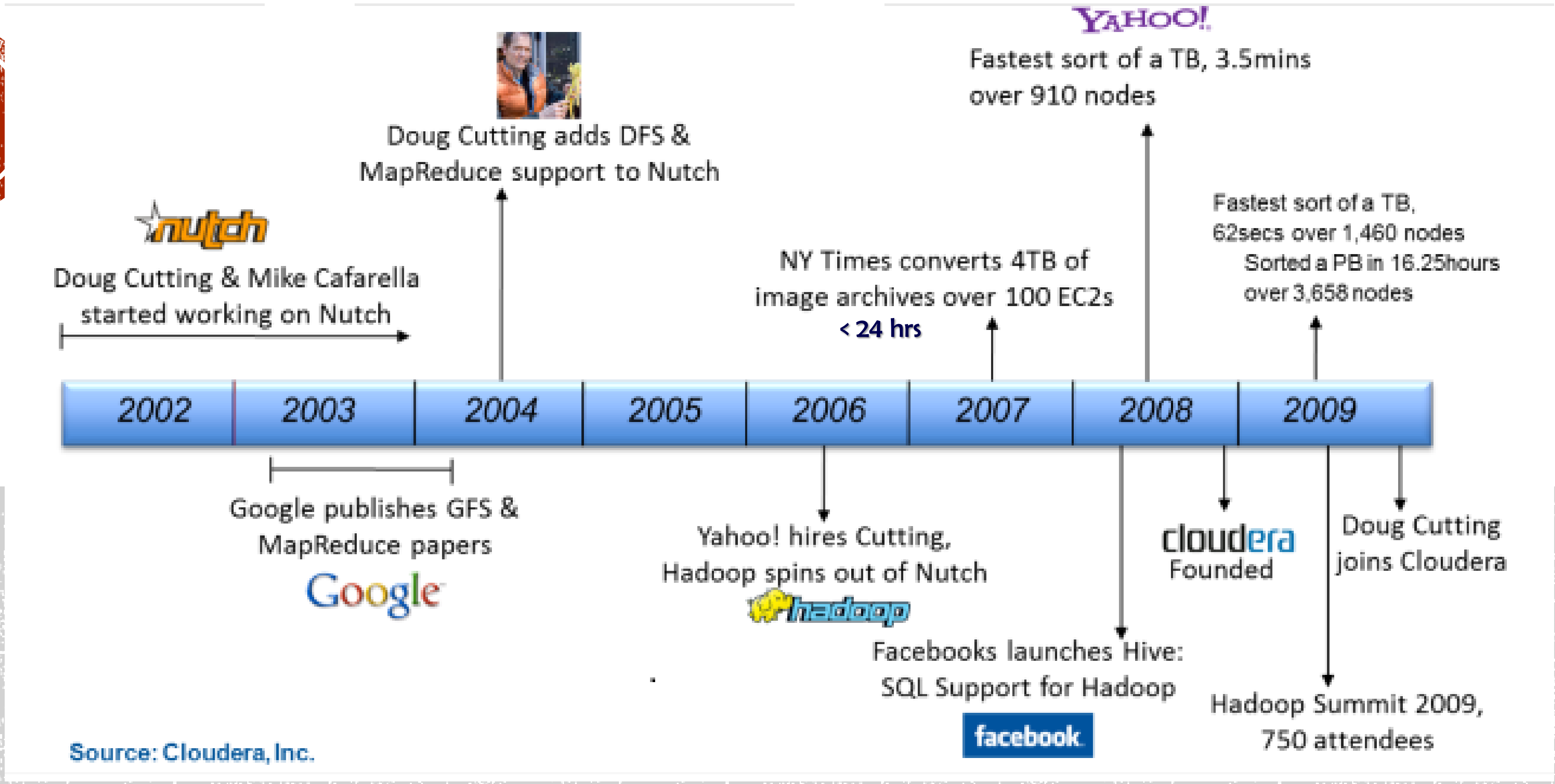
HADOOP HISTORY

Apache Lucene project – text search library

Apache Nutch – open source web search engine for Lucene

- Index and crawl → need big cluster to process/ expensive to invest
- One billion pages index cost $500k in hardware + $30k per month

In 2008, Hadoop became the Apache top level project



2002 2003 2004 2005 2006 2007 2008 2009

Doug Cutting & Mike Cafarella started working on Nutch

Google publishes GFS & MapReduce papers

Doug Cutting adds DFS & MapReduce support to Nutch

Yahoo! hires Cutting, Hadoop spins out of Nutch

NY Times converts 4TB of image archives over 100 EC2s

< 24 hrs

Facebooks launches Hive: SQL Support for Hadoop

YAHOO!

Fastest sort of a TB, 3.5mins over 910 nodes

cloudera Founded

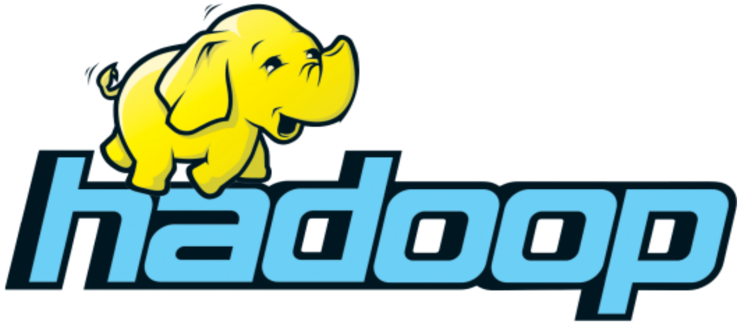
Fastest sort of a TB, 62secs over 1,460 nodes

Sorted a PB in 16.25hours over 3,658 nodes

Hadoop Summit 2009, 750 attendees

Doug Cutting joins Cloudera

Source: Cloudera, Inc.



Where does the name come from?

"The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria." … Doug Cutting

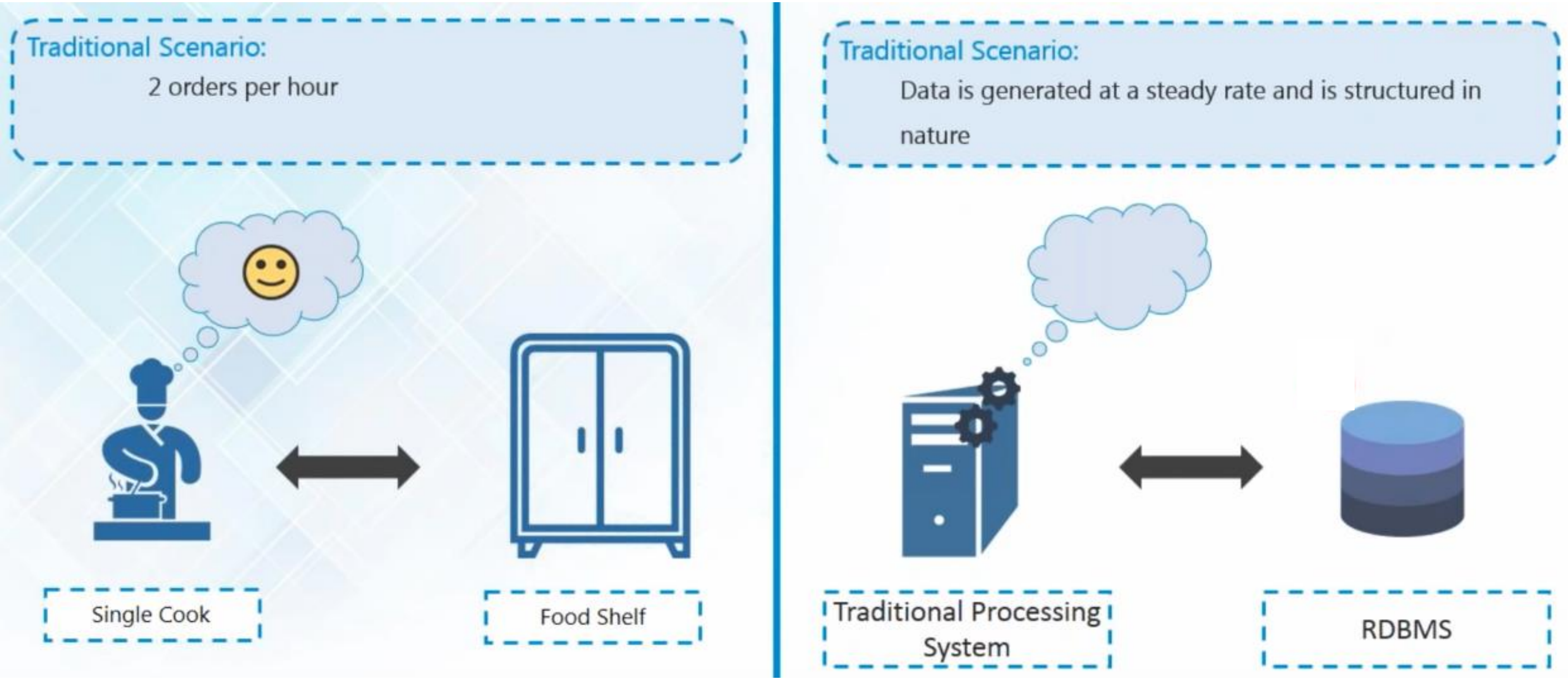
HADOOP MAIN COMPONENTS

- ***Hadoop Distributed File System (HDFS)***
  - Designed to provide highly fault-tolerant and to be deployed on low-cost hardware

- ***MapReduce***
  - A framework for processing data in batch - BSP

WHY HADOOP?

SMALL SCALE SYSTEM



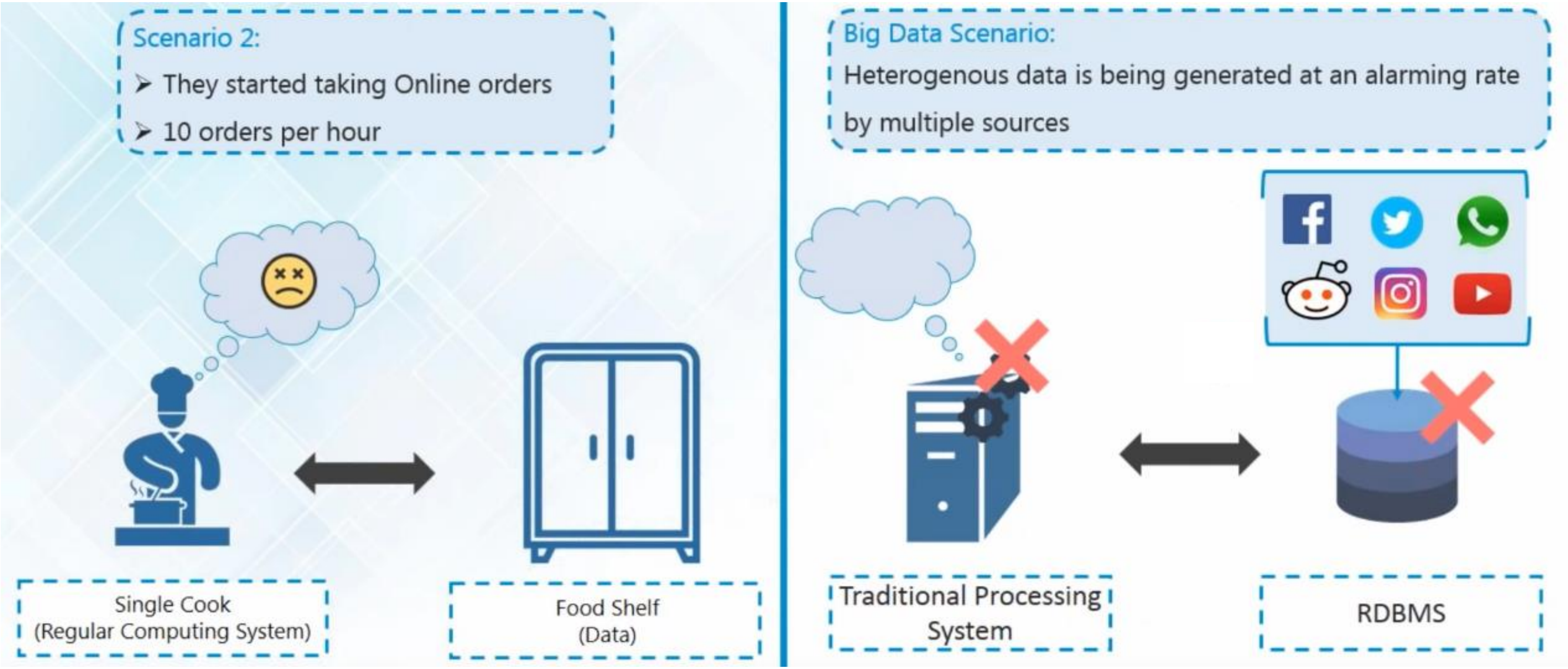
Traditional Scenario:

2 orders per hour

Traditional Scenario:

Data is generated at a steady rate and is structured in nature

BUT …



Scenario 2:

- ➢ They started taking Online orders
- ➢ 10 orders per hour

Single Cook
(Regular Computing System)

Food Shelf
(Data)

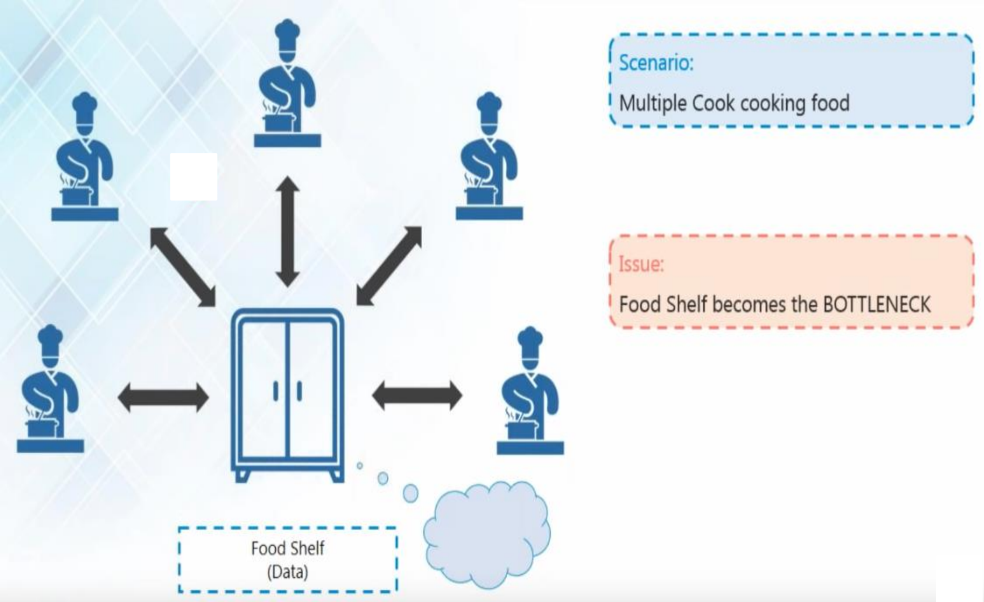
Big Data Scenario:

Heterogenous data is being generated at an alarming rate by multiple sources

Traditional Processing System

RDBMS

SOLVING THE PROBLEM – PHASE I



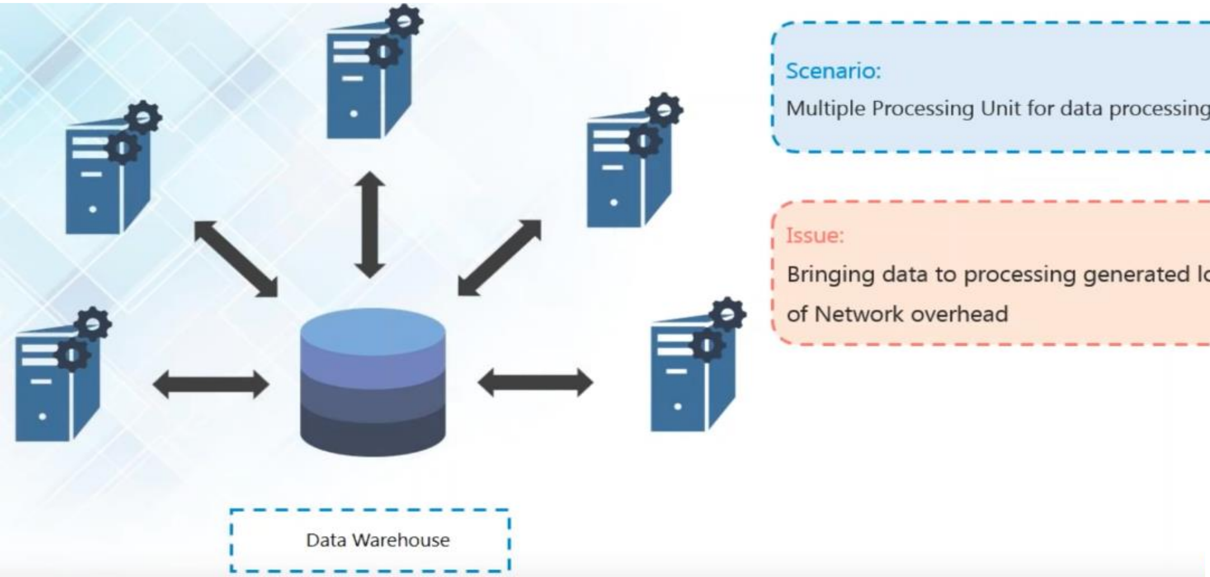
Scenario:

Multiple Cook cooking food

Issue:

Food Shelf becomes the BOTTLENECK

Food Shelf
(Data)



Scenario:

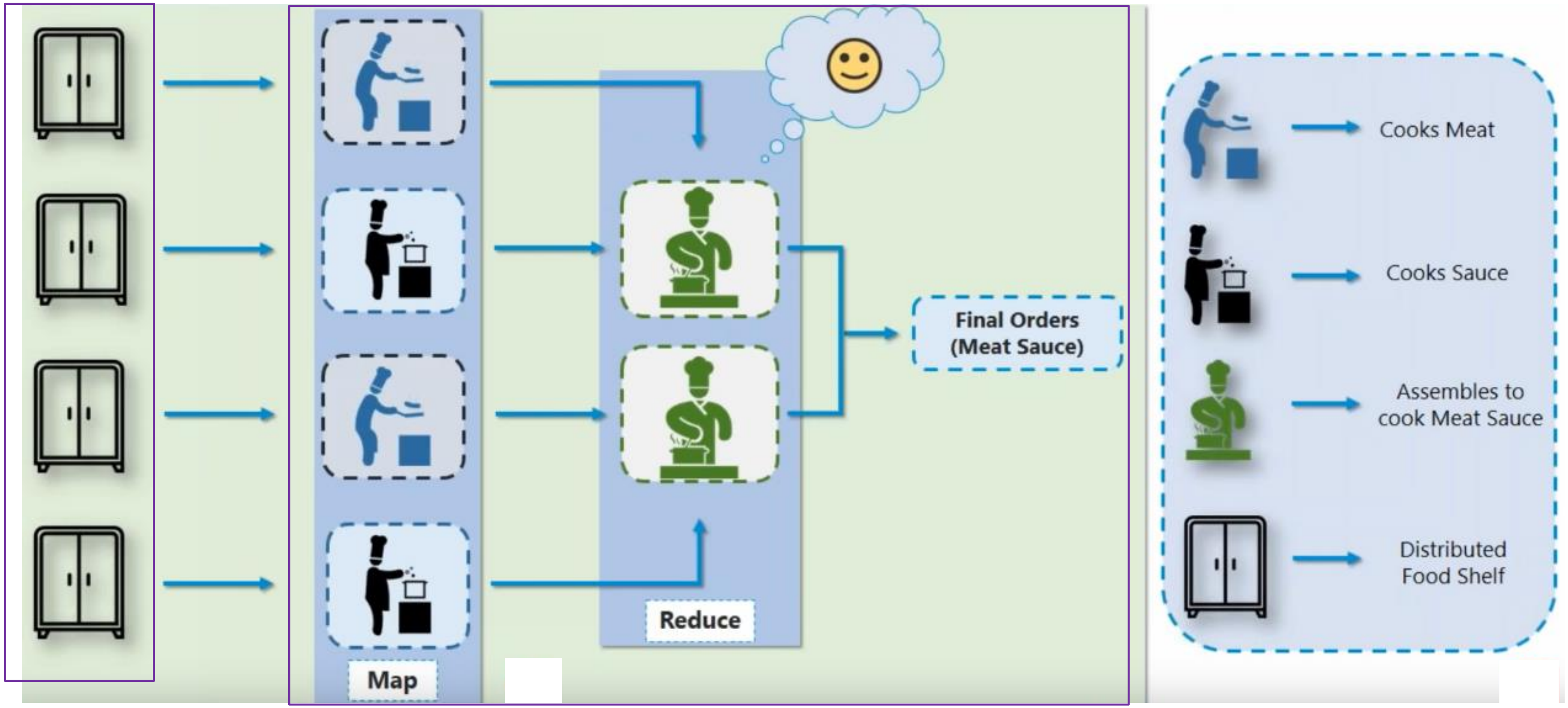
Multiple Processing Unit for data processing

Issue:

Bringing data to processing generated lots of Network overhead

Data Warehouse

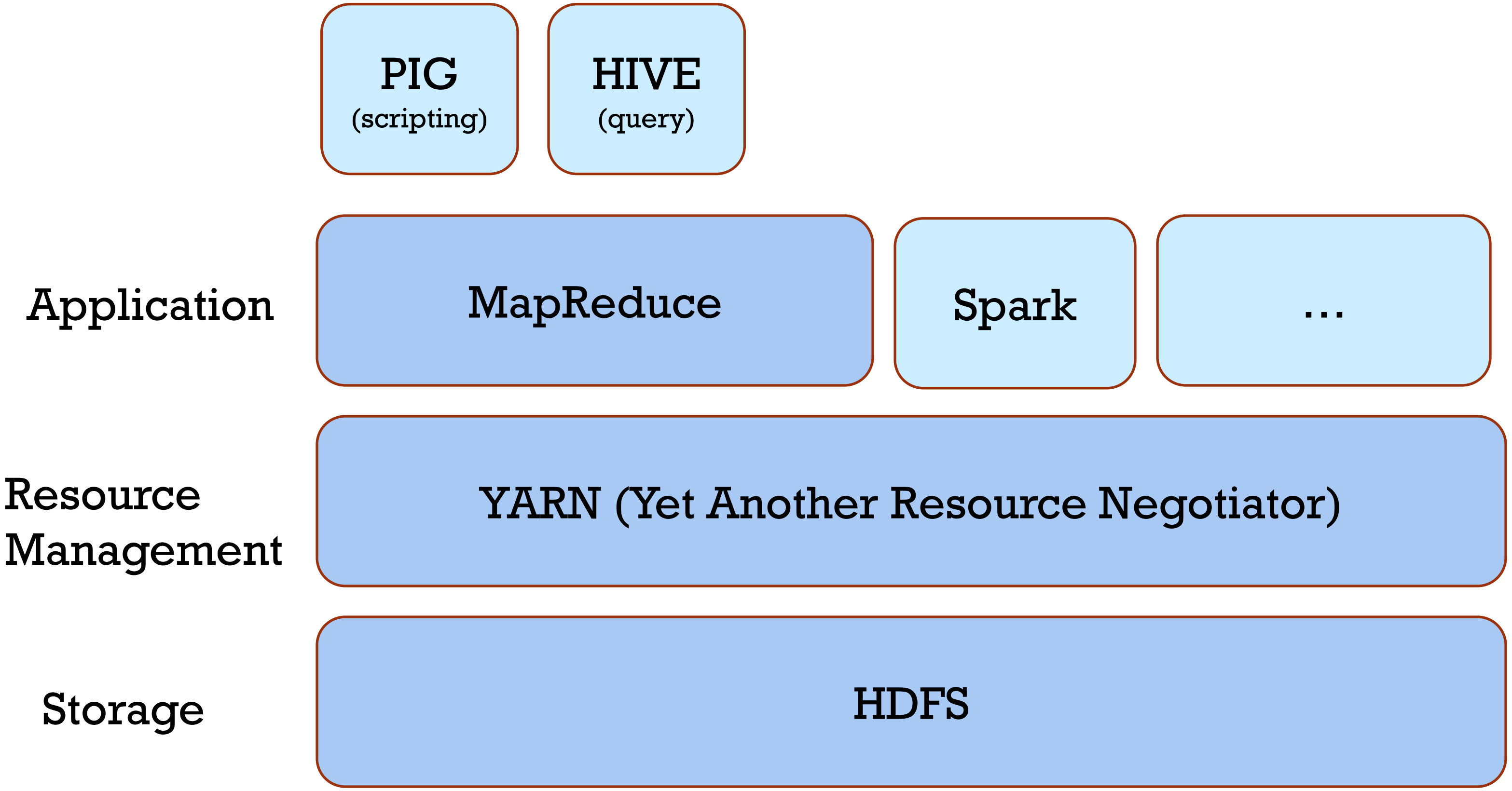
SOLVING THE PROBLEM – PHASE II



HDFS

Map/Reduce

HADOOP DETAILED SOFTWARE ARCHITECTURE



PIG
(scripting)
HIVE
(query)
Application
MapReduce
Spark
…
Resource Management
YARN (Yet Another Resource Negotiator)
Storage
HDFS