Introduction to Big Data

ECAP456

Dr. Rajni Bhalla
Associate Professor

Learning Outcomes

After this lecture, you will be able to

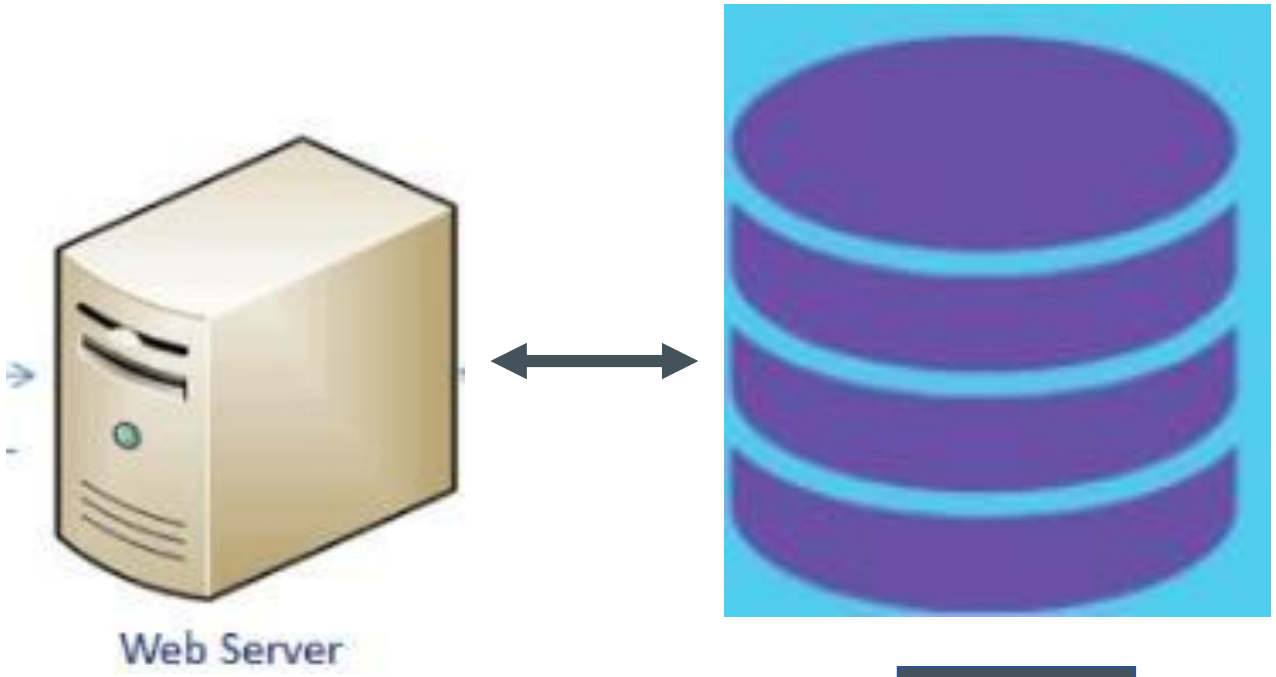
- Learn introduction about Hadoop.
- Learn how Hadoop Improves on Traditional Databases
- Understand Why is Hadoop important?
- Understand what are the challenges of using Hadoop?
- Learn benefits of Hadoop for bigdata

Introduction

Introduction

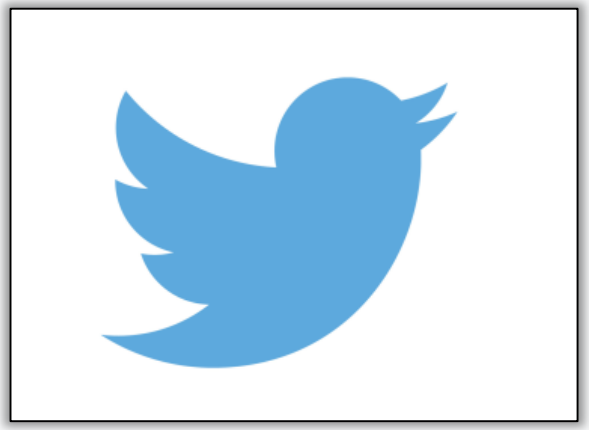
Traditional Scenario

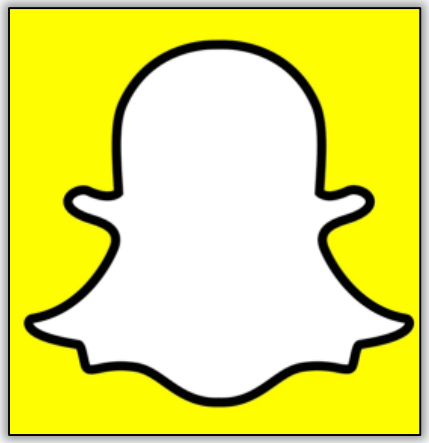
Data is generated at a steady rate and is structured in nature.



Introduction

Introduction



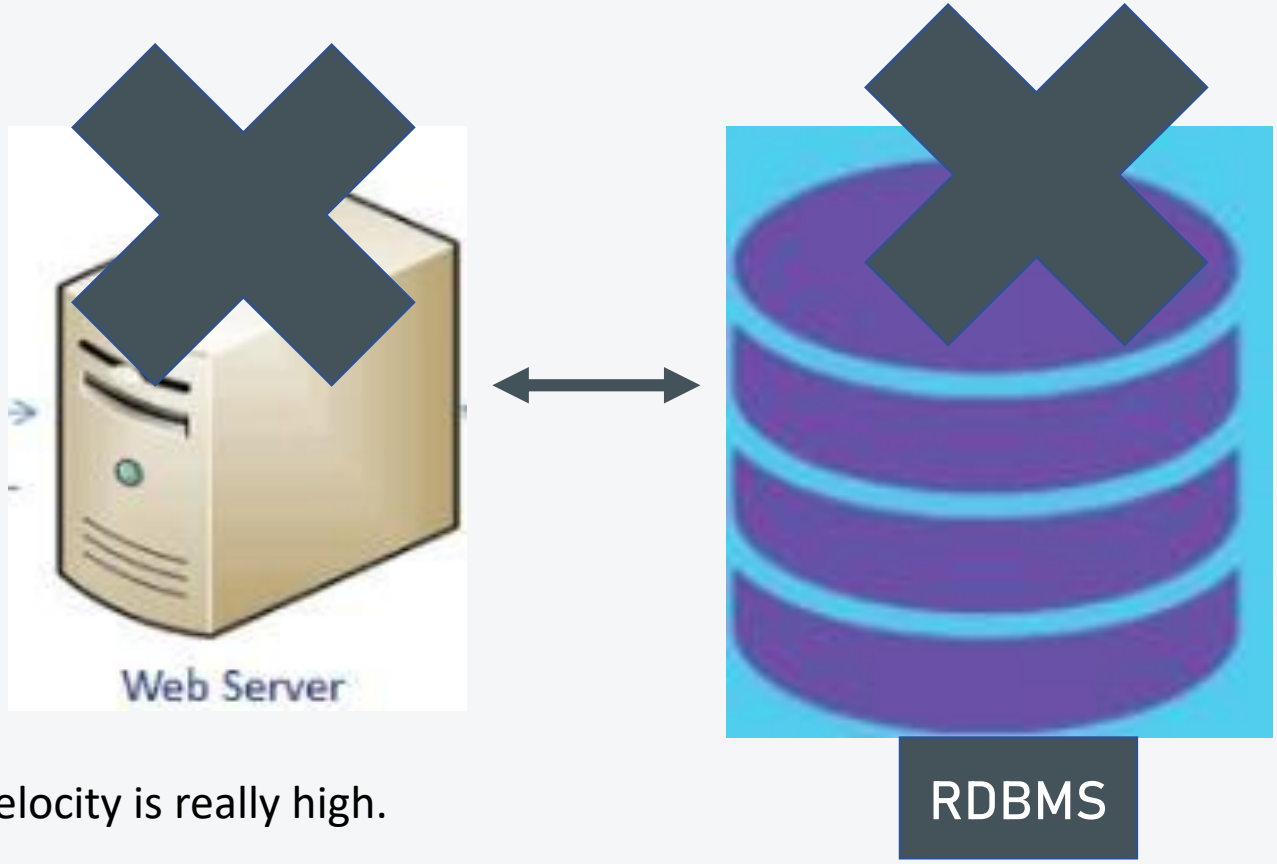


Failure of traditional system

Receiving much more order than expected
Cook is not capable of cooking 10 dishes per hour.

Failure of Traditional Scenario

Heterogeneous data is being generated at an alarming stage by multiple sources



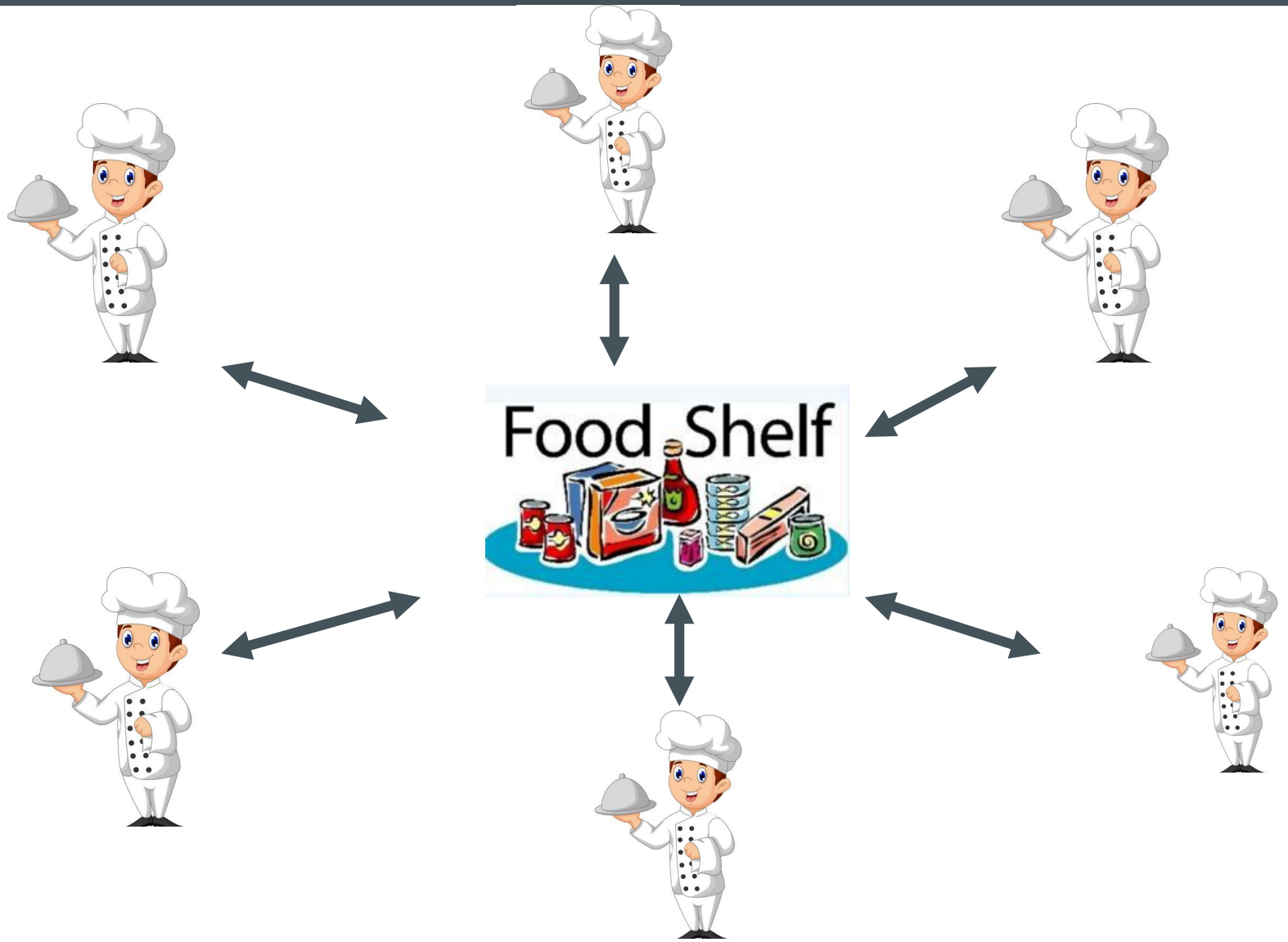
Velocity is really high.

Failure of Traditional Scenario

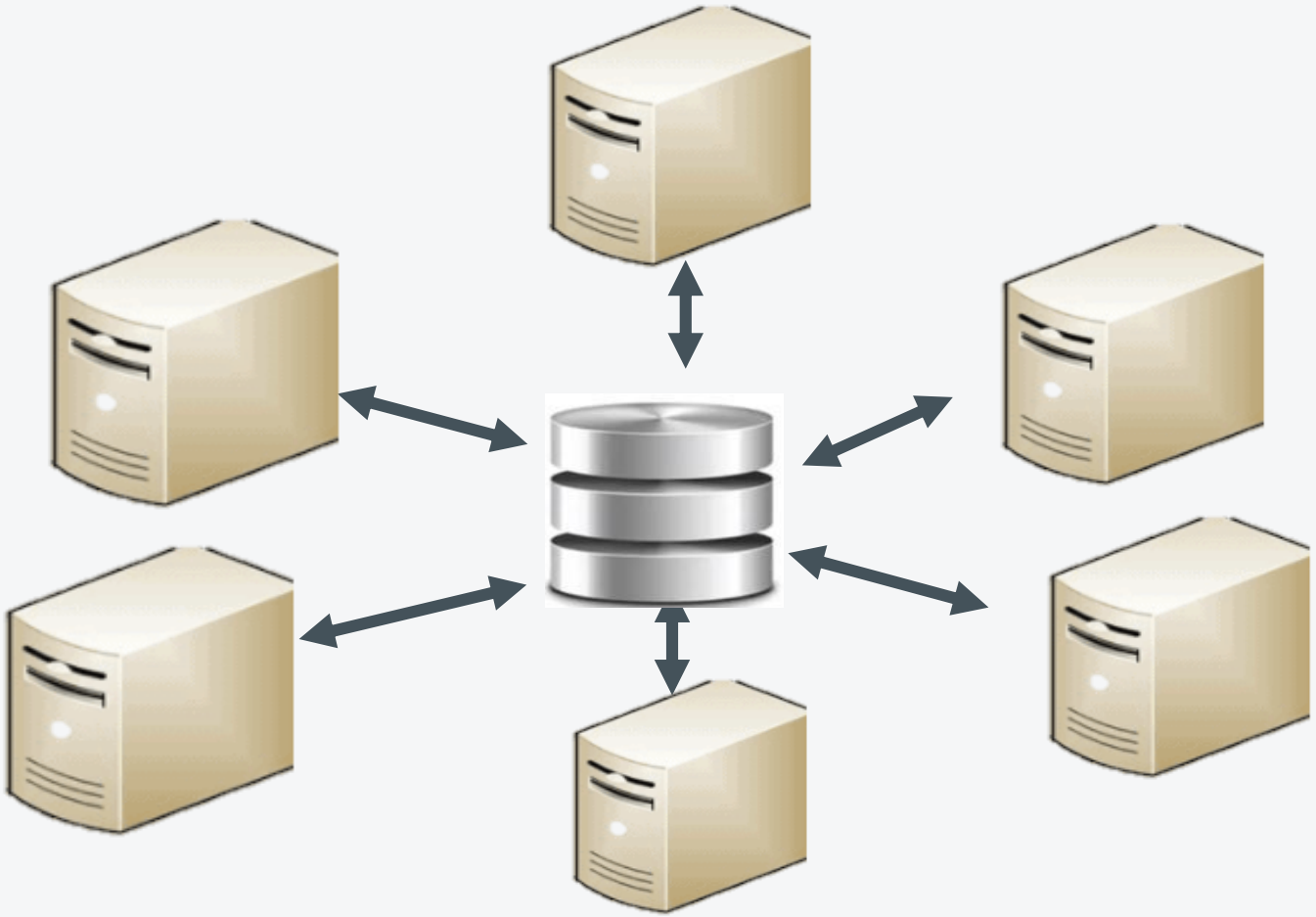
ISSUE1: TOO MANY ORDERS PER HOUR

SOLUTION: HIRE MULTIPLE COOK

Introduction



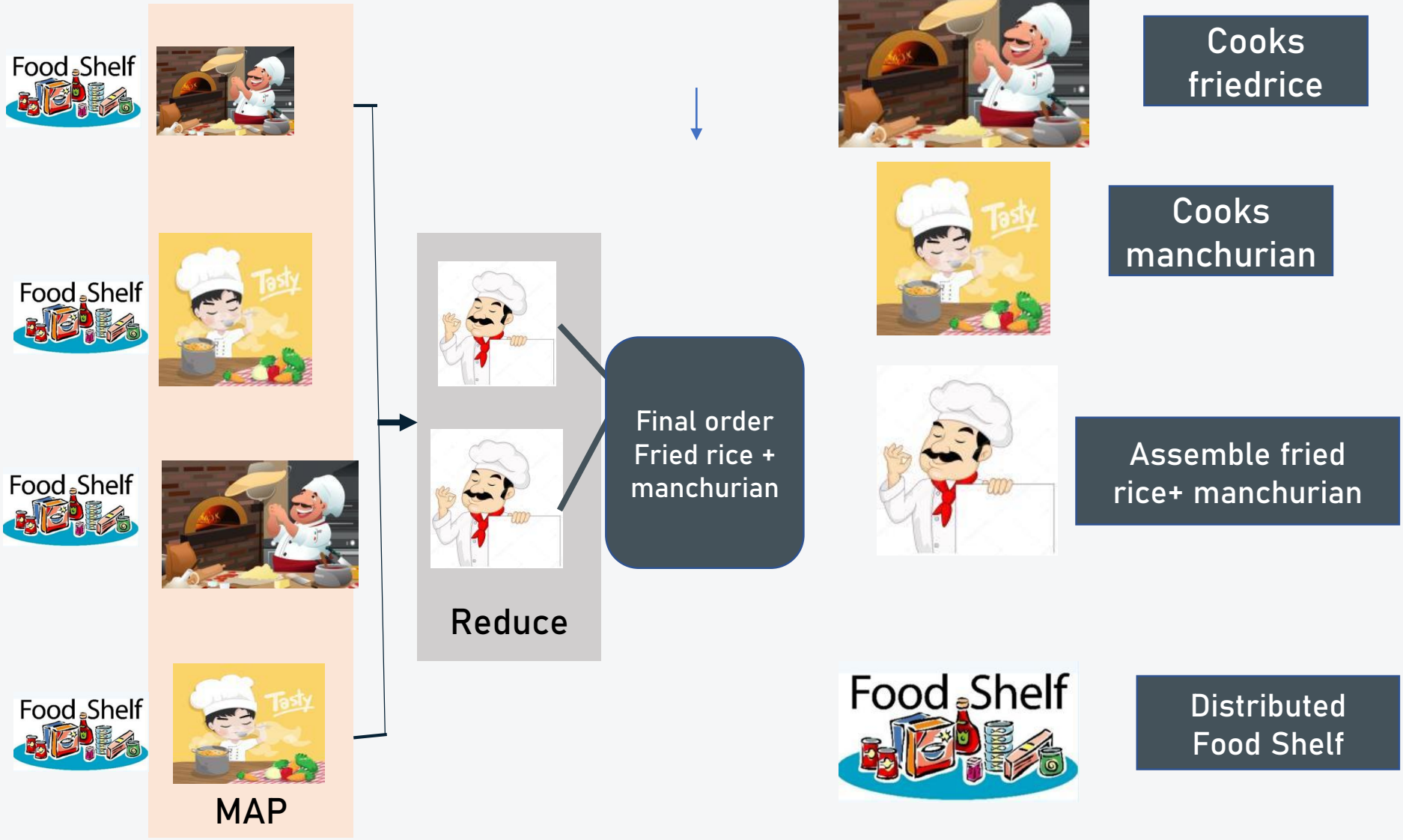
Multipleprocessing unit for processing



ISSUE: FOODSHELF BECOME THE BOTTLENECK

SOLUTION: DISTRIBUTED AND PARALLEL PROCESSING

Introduction



Do we have a framework that works like that for storing and processing big data?

Framework to process BIG DATA

Introduction

Hadoop is a framework that allows us to store and process large datasets in parallel and distributed fashion.

Hadoop

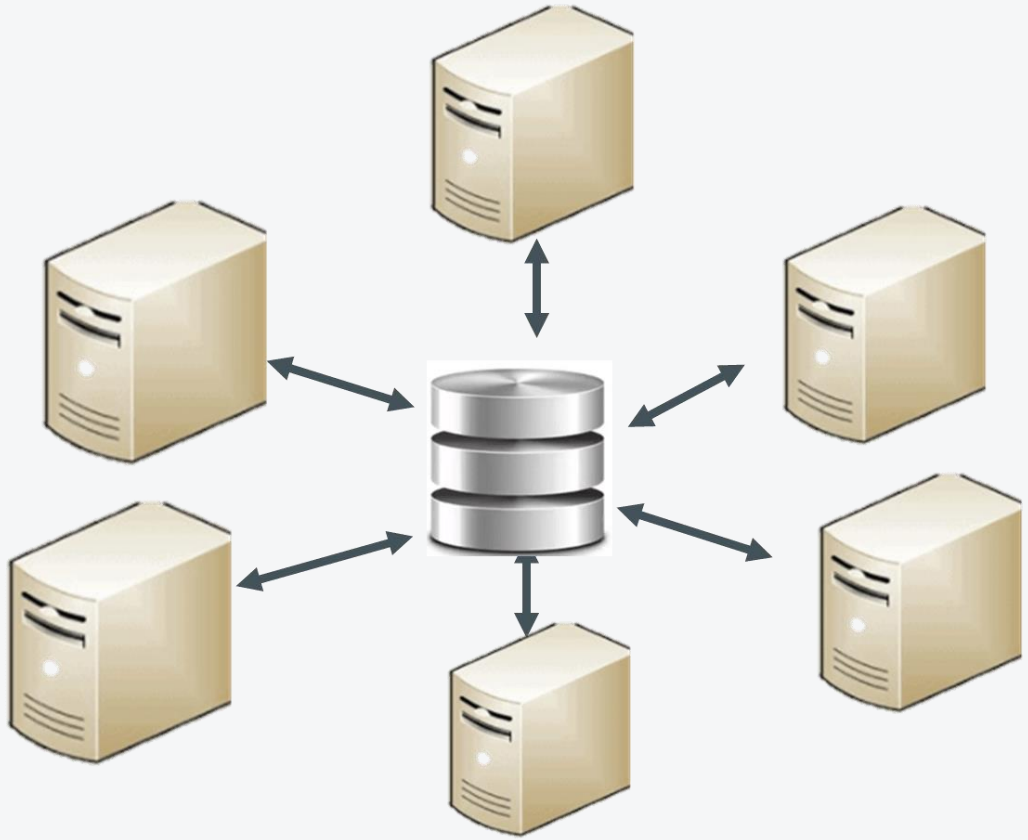
Two major problems in dealing with BIG DATA

- Storage
- Processing

Hadoop

Storage problem resolved by

- HDFS



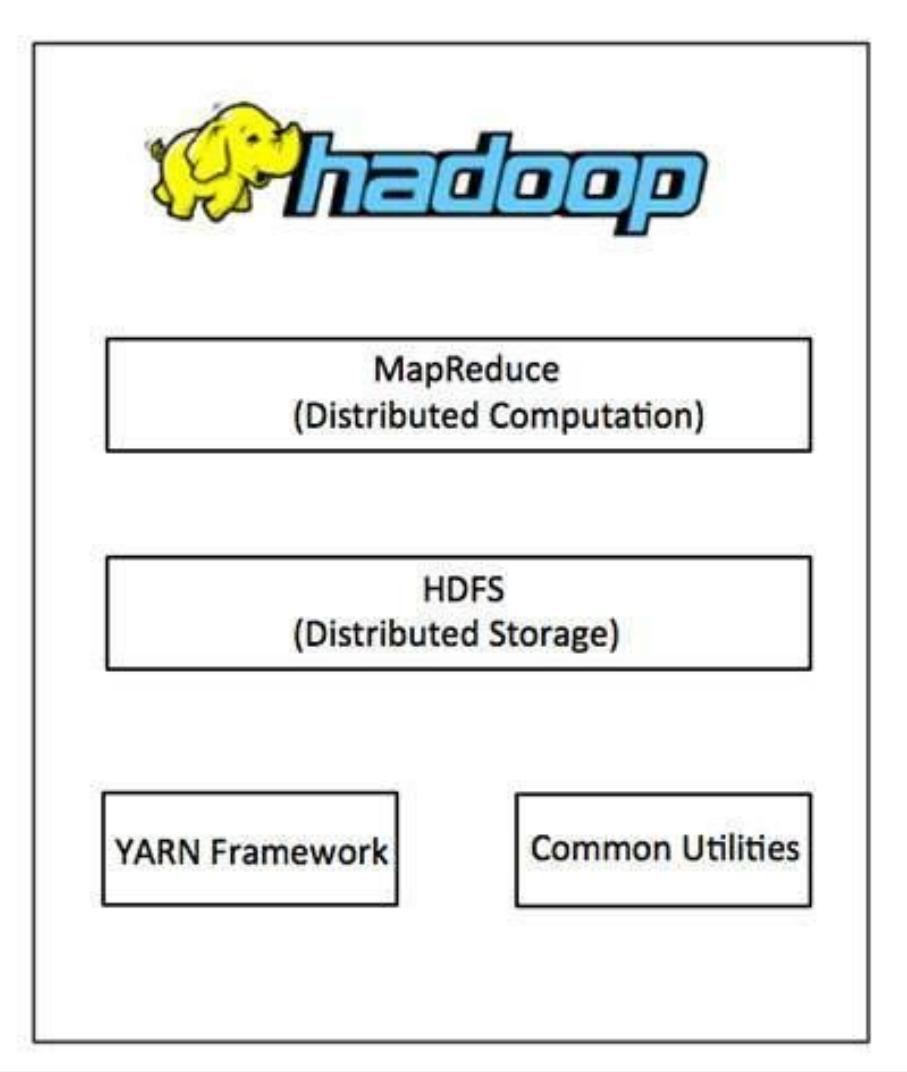
Distributed File System

Hadoop

- Processing problem resoved by
  - mapReduce

Processing:Allow distributed and parallel processing

Hadoop Architecture



Hadoop Architecture

At its core, Hadoop has two major layers namely

- Processing/Computation layer (MapReduce), and
- Storage layer (Hadoop Distributed File System)

How Hadoop Improves on Traditional Databases

- Capacity: Hadoop stores large volumes of data.

How Hadoop Improves on Traditional Databases

- Capacity: Hadoop stores large volumes of data.
- Speed: Hadoop stores and retrieves data faster.

Why is Hadoop important?

- Ability to store and process huge amounts of any kind of data, quickly.
- Computing power.
- Fault tolerance.
- Flexibility.
- Low cost.
- Scalability.

What are the challenges of using Hadoop?

- MapReduce programming is not a good match for all problems.
- There's a widely acknowledged talent gap.
- Data security.
- Full-fledged data management and governance

Benefits of Hadoop for Big Data

1. Resilience
2. Scalability
3. Low cost
4. Speed
5. Data diversity

The Hadoop Ecosystem: Supplementary Components

- HDFS: Hadoop Distributed File System
- YARN: Yet Another Resource Negotiat or
- MapReduce: Programming based Data Processing
- Spark: In-Memory data processing
- PIG,HIVE: Query based processing of data services

The Hadoop Ecosystem: Supplementary Components

- HBase: NoSQL Database
- Mahout,Spark MLLib: Machine Learning algorithm librarie
- Solar,Lucene: Searching and Indexing
- Zookeeper: Managing cluster
- Oozie: Job Scheduling


That's all for now...