**1. Big Data?**

Traditional dilemma: Too many request -> multiple processing unit for data processing

New dilemma: Data warehouse ->

A screenshot of a cell phone

Description automatically generated

Restaurant example: distributed the shelf and each chief gets its own shelf

**2. Apache Hadoop: Framework to Process Big Data**

A screenshot of a cell phone

Description automatically generated

Hadoop Master/ Slave Architecture:

Example: project manager-> worker1, worker2, worker3, worker4

Each task has backup

A picture containing text

Description automatically generated

**3. HDFS**

NameNode & DataNode

A close up of a map

Description automatically generated

NameNode:

* Maintains and manages DataNodes
* Records metadata i.e. information about data blocks e.g. location of blocks stored, the size of the files, permissions, hierarchy, etc.
* Receives heartbeat and block report from all the DataNodes

DataNode:

* Slave daemons
* Stores actual data
* Serves read and write requests from the client

A close up of a map

Description automatically generated

fsImage: all data changes since the NameNode start (on disk)

editLog: data changes in last 1 hour (in RAM)

**3.1 HDFS Data Blocks—How the data is actually stored in DataNodes**