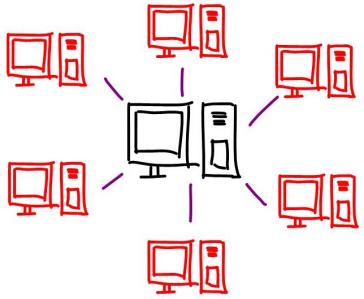
Azure Hadoop

Eshant Garg
Advisor, Data Specialist
Eshant.garg@gmail.com

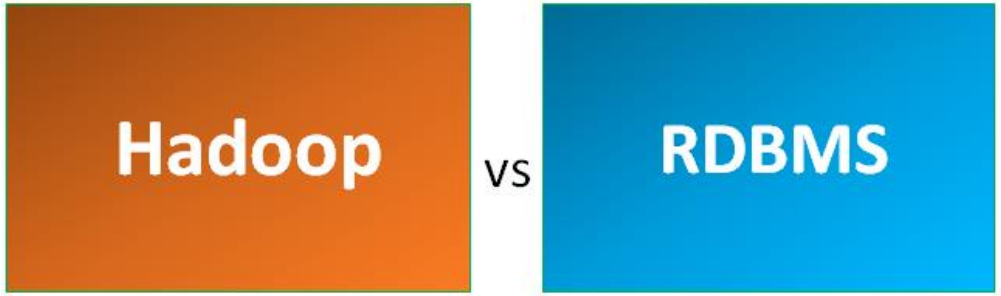
Module Overview

Why Traditional Systems are failing?



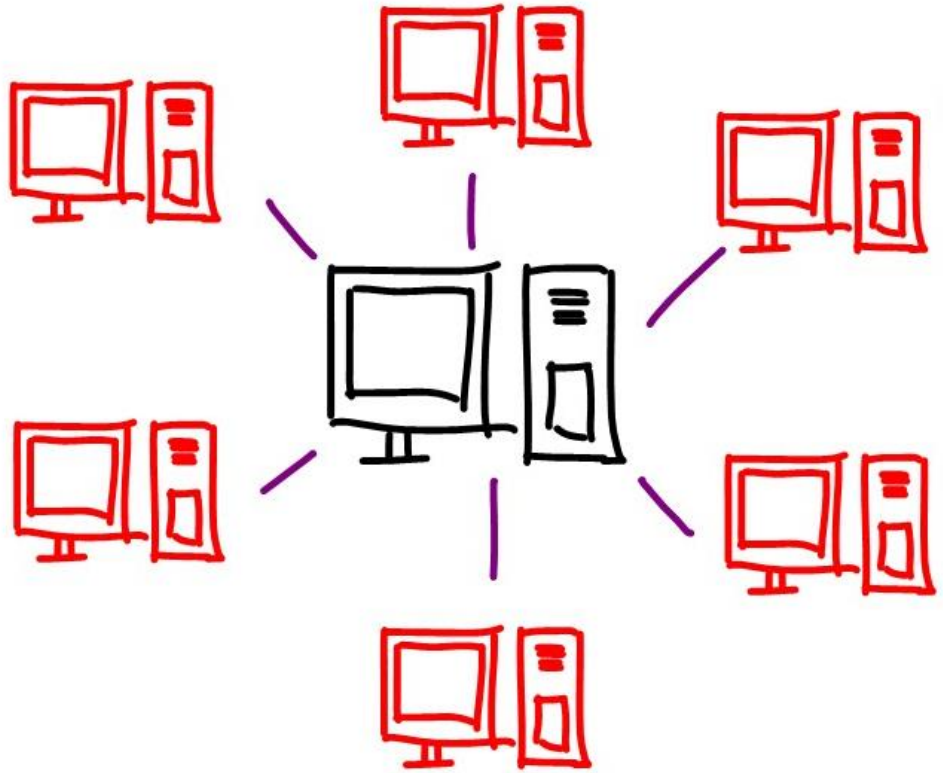
Why Distributed Computing System?

Introducing Hadoop?



Hadoop Ecosystem?

Need of Distributed Computing?

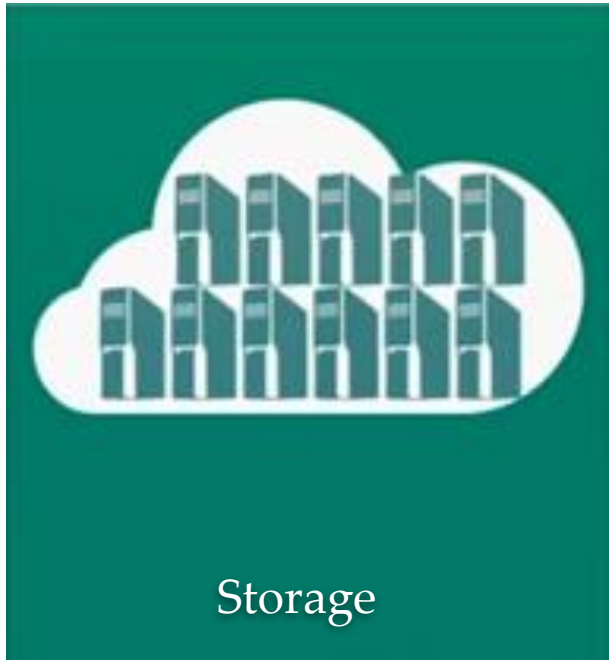


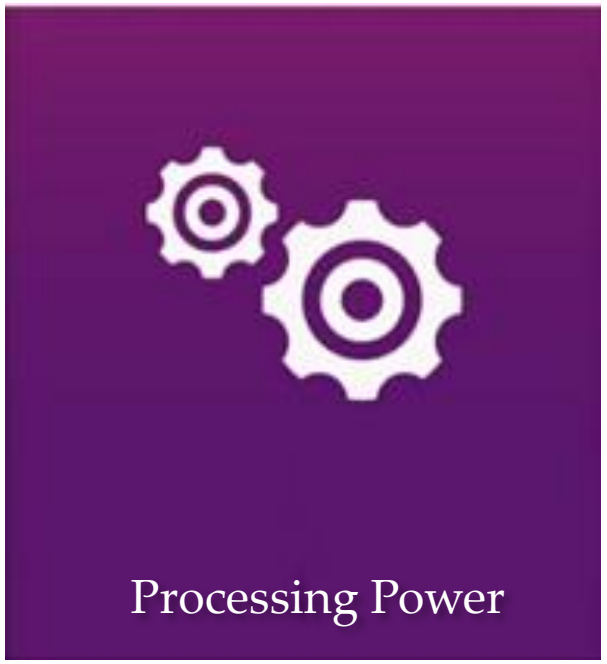
How much data?

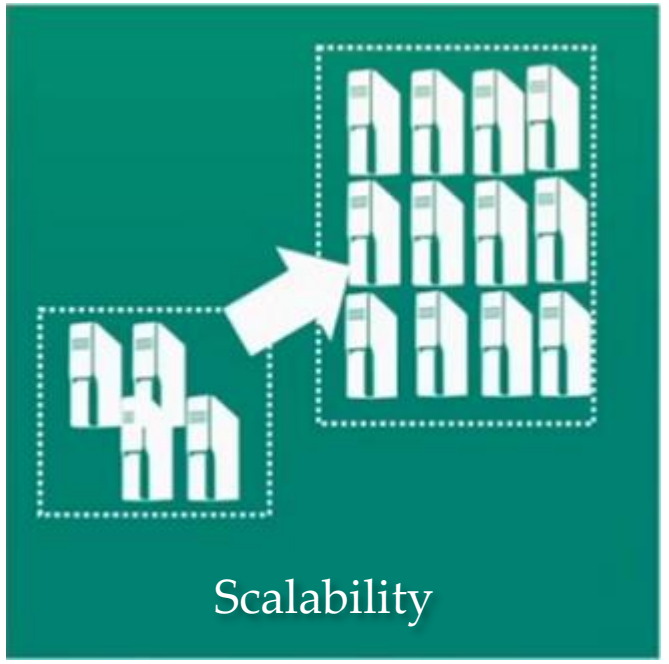
- 2.4 billion monthly active users
- Generate 4 petabytes of data every single day
- 100 million hours of video watch time per hour
- 4 million like every minute

- Stores 20 EB of data
- 4 million searches happening every minute
- 4 million apps on google play
- 300 hours of video upload every minute

Requirement to handle Big Data?





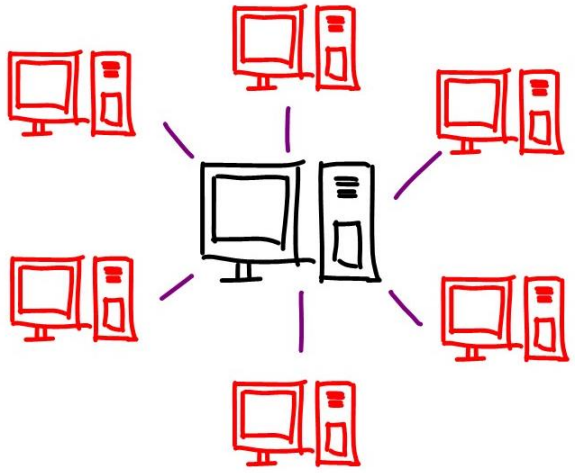


Monolithic system

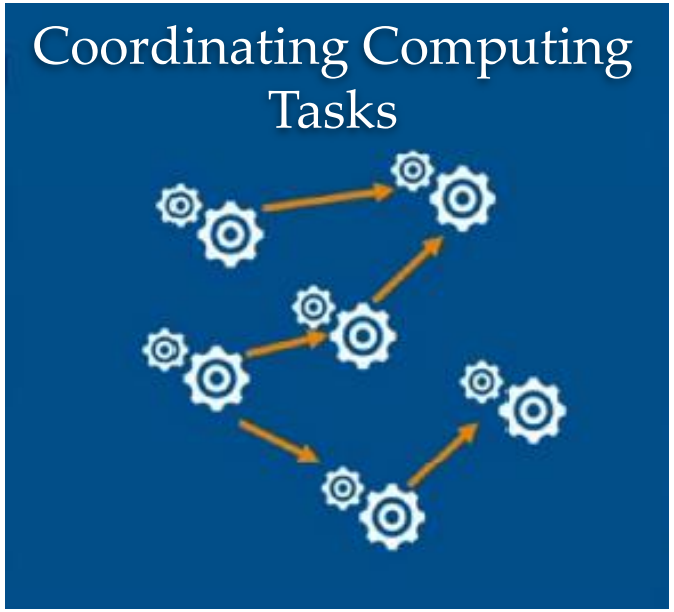
- Single machine
- Single process
- Powerful single server
- Can not scale beyond limit

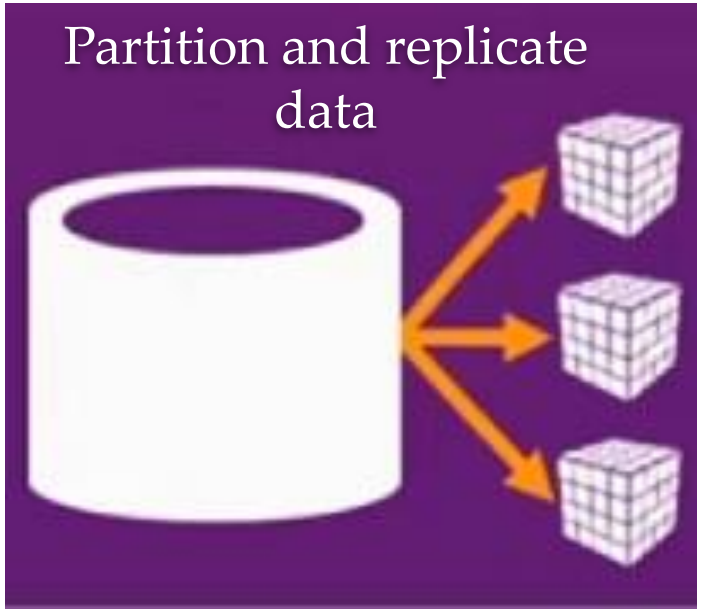
Distributed System

- Cluster of multiple machines
- Multiple processes
- Commodity hardware
- Can scale storage and computational capacity linearly



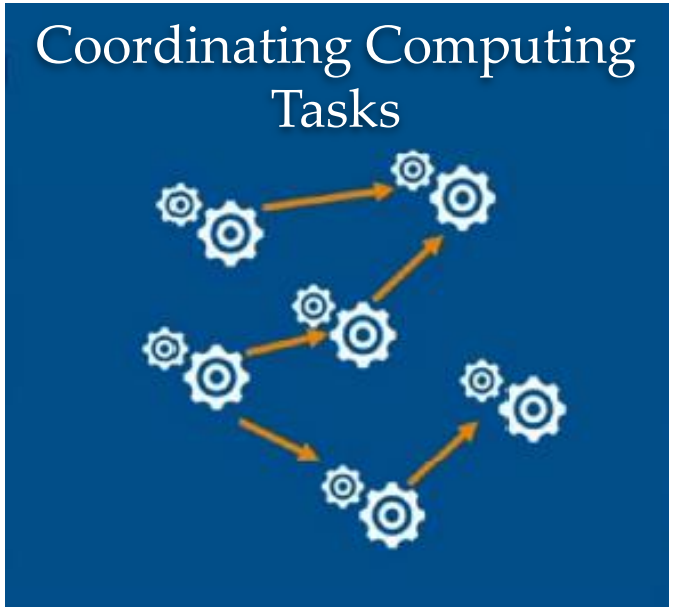
Software requirements to handle Distributed systems?

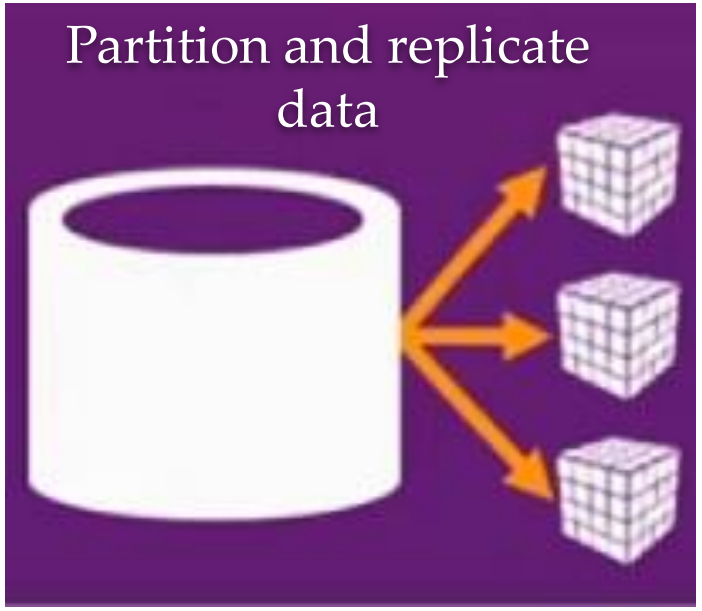




Software purpose is to coordinate and manage all the processes and machines which exist within the system.

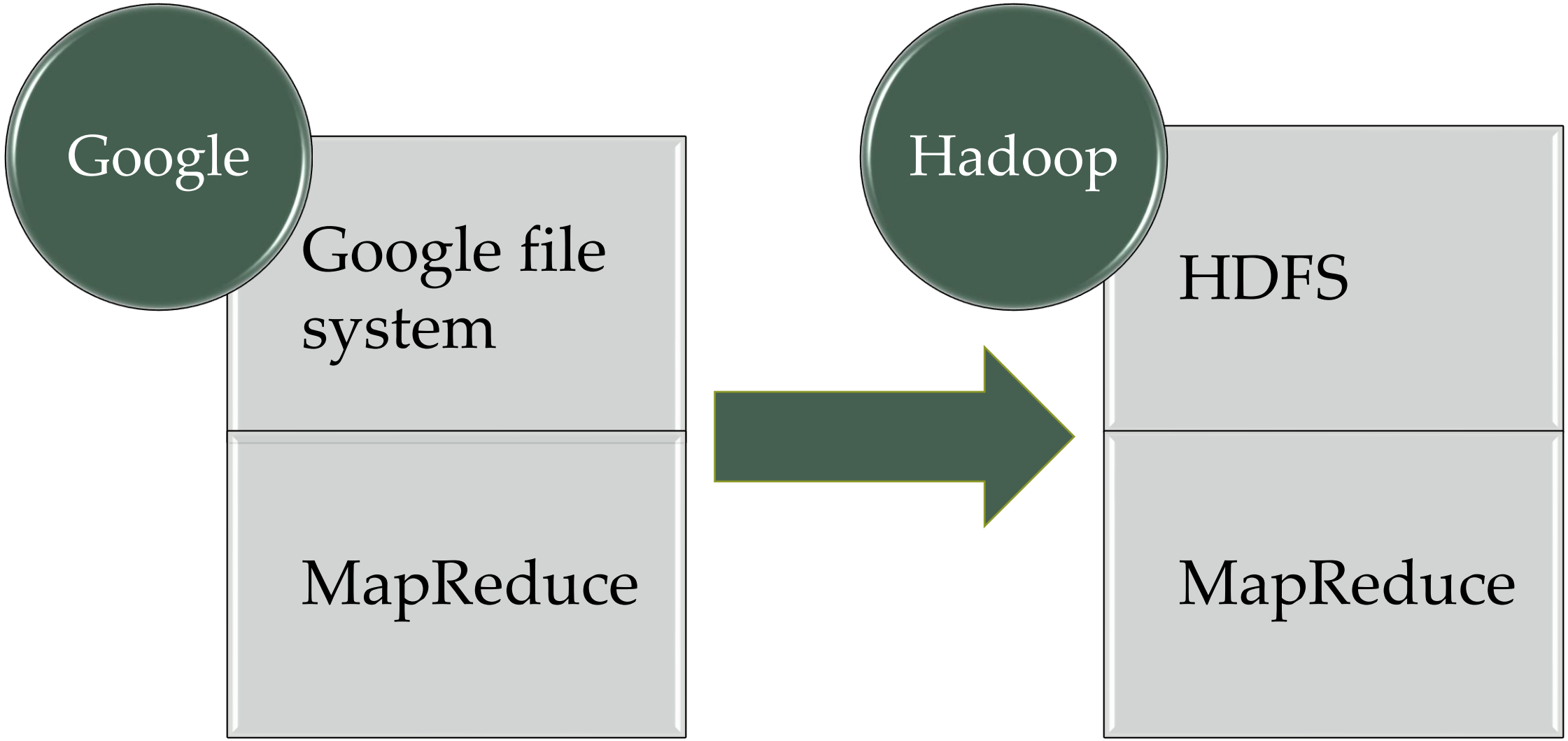
Google Software Challenge?



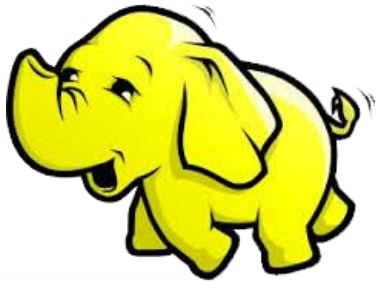


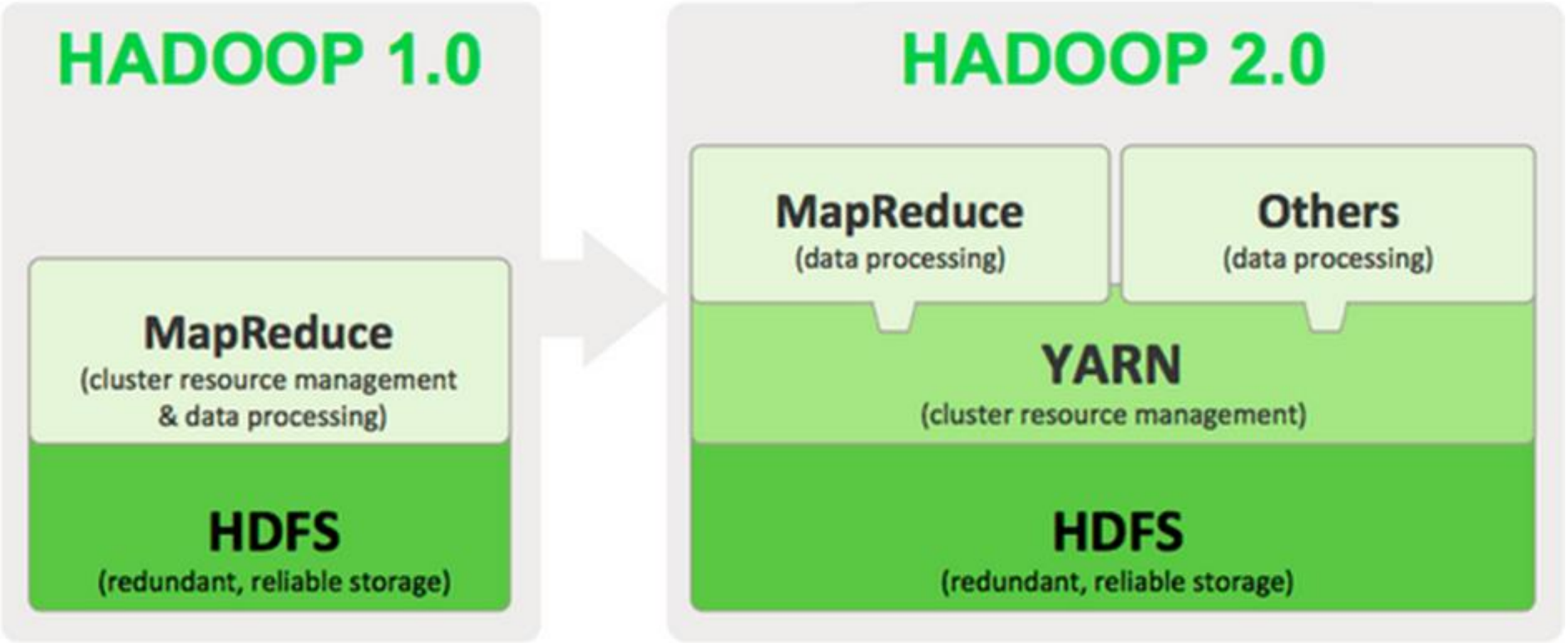
Google software challenge?

- Can store millions of records across multiple machines → Google file system
- Can run and coordinate these processes across all these machines → MapReduce

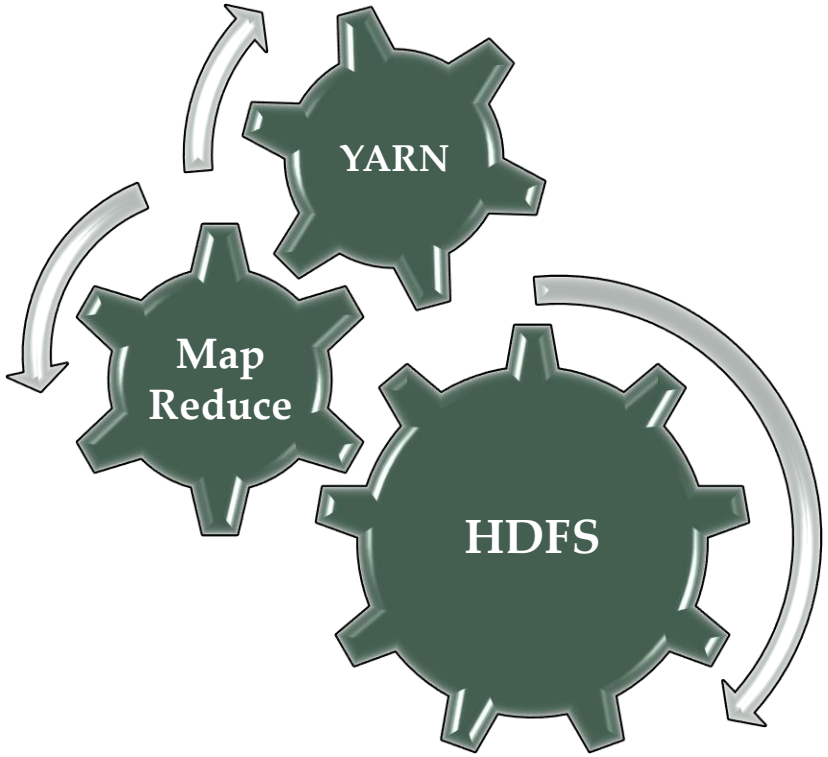

Google
Google file system
MapReduce
Hadoop
HDFS
MapReduce

Hadoop Architecture

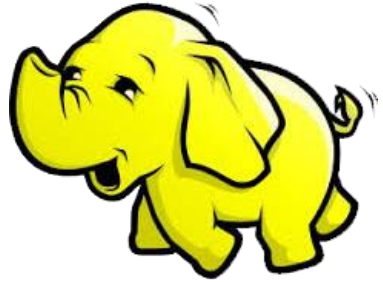


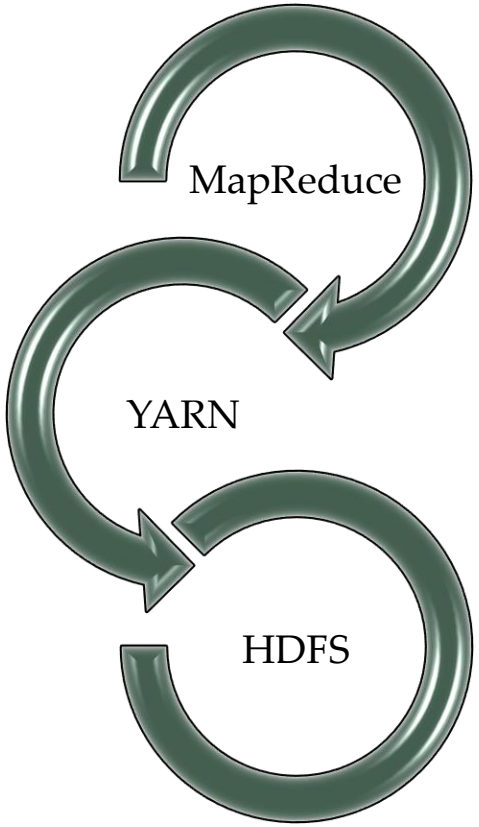


Hadoop Architecture



What happens when you submit a job Hadoop?





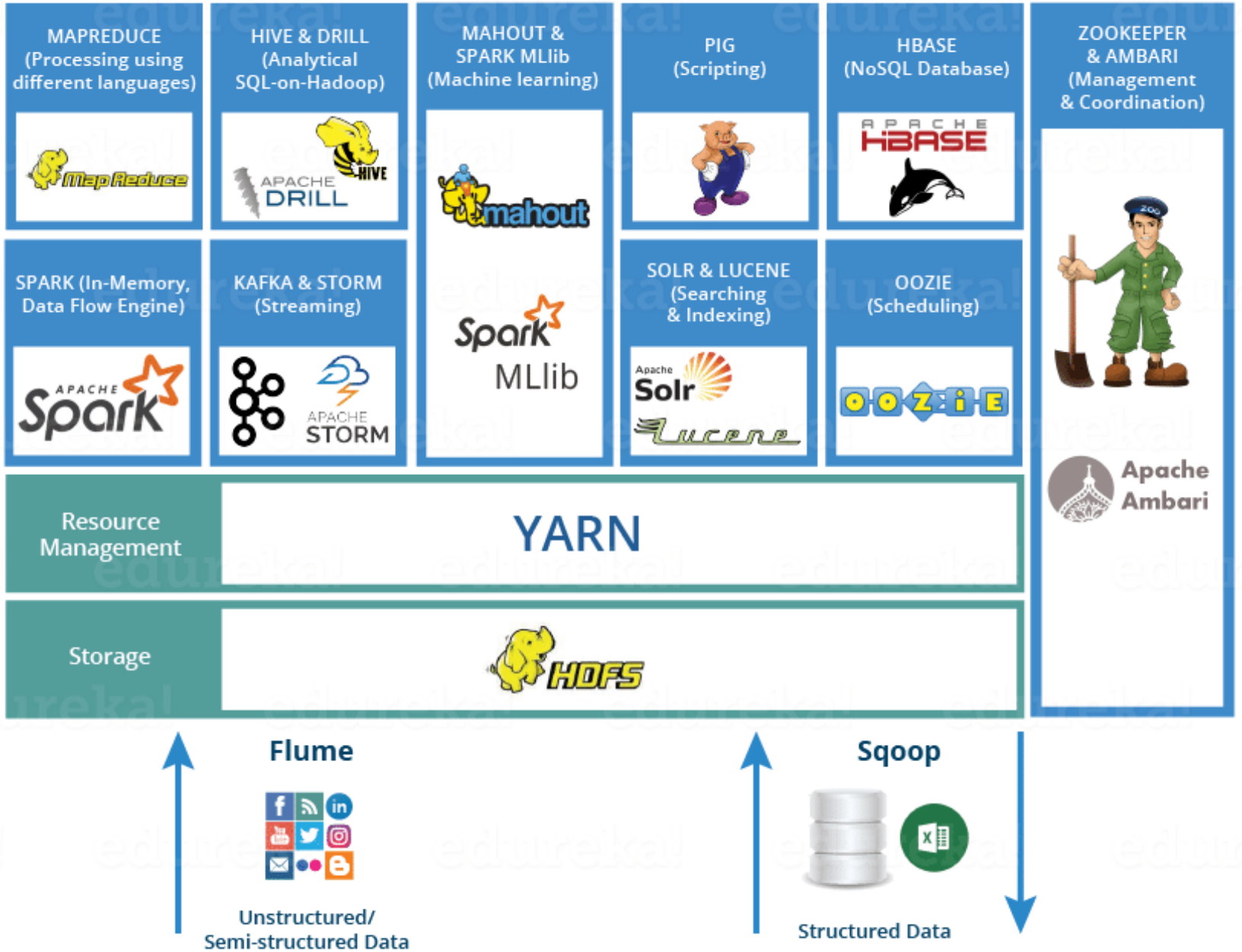
Hadoop vs RDBMS

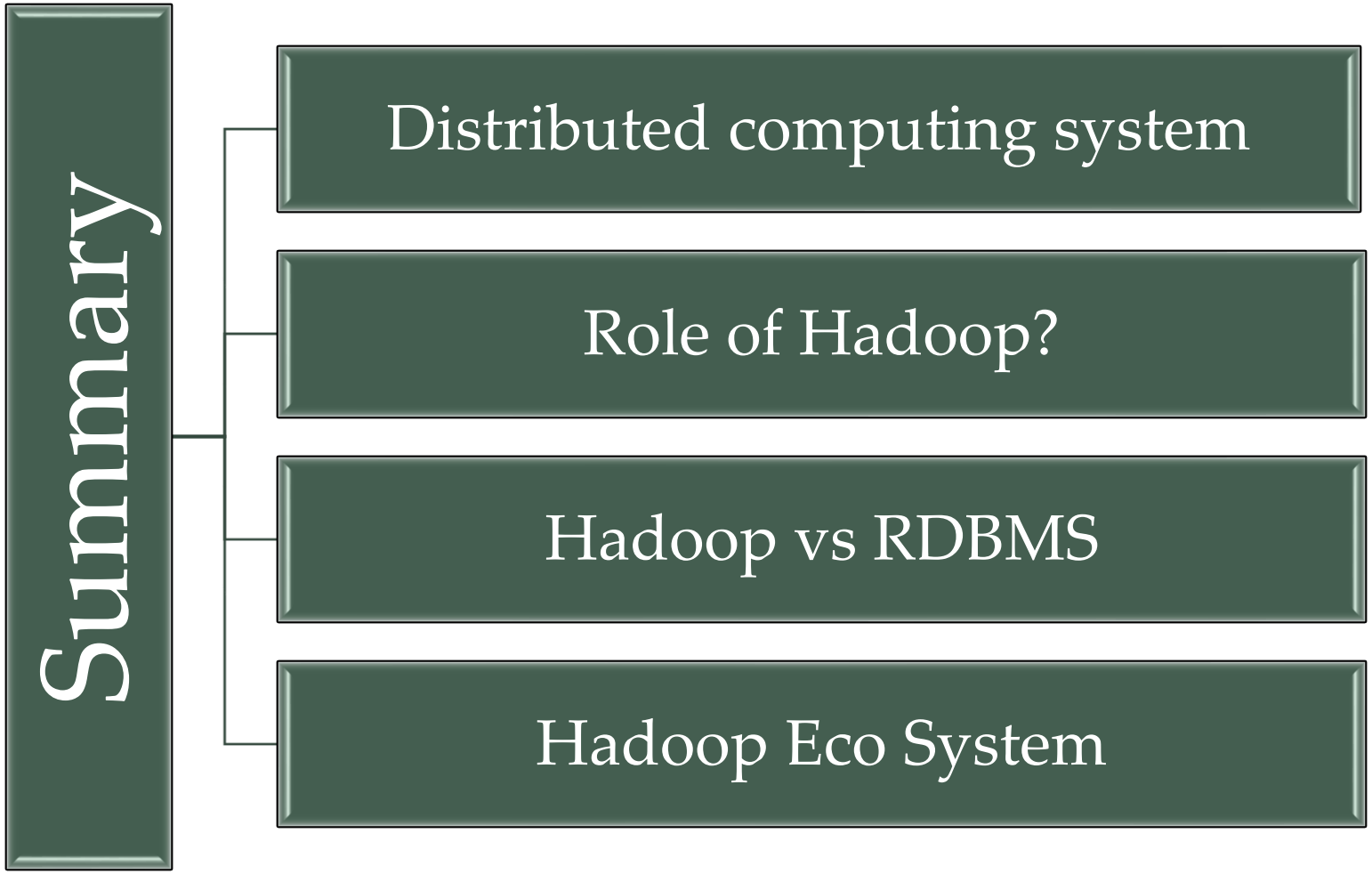
Hadoop

- Unstructured
- CAP
- Higher data throughput
- Slower granular query performance
- Horizontally scaled
- OLAP

RDBMS

- Structured
- ACID
- Lower data throughput
- Faster granular query performance
- Vertically scaled
- OLTP

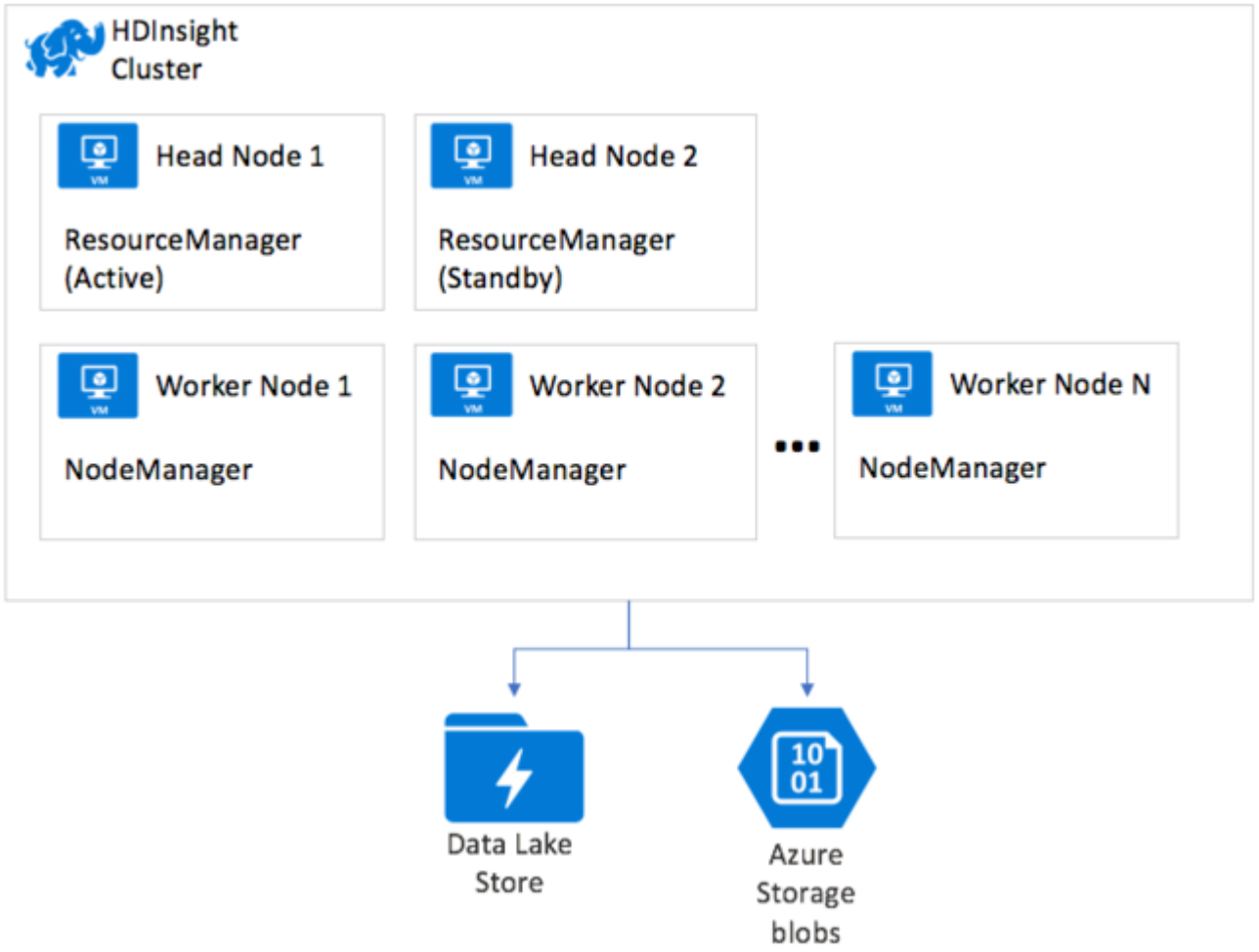

MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & Indexing)
OOZIE
(Scheduling)
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/
Semi-structured Data
Sqoop
Structured Data


Summary
Distributed computing system
Role of Hadoop?
Hadoop vs RDBMS
Hadoop Eco System

HDInsight high level architecture

Parallel Processing

Decoupled Storage



MapReduce operation

Data is chunked redundantly across nodes

Massive Parallelism