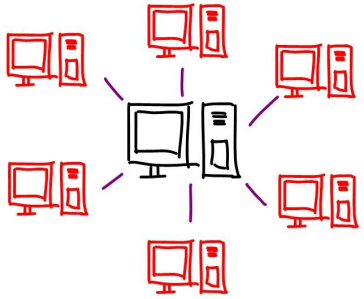
Azure Hadoop

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

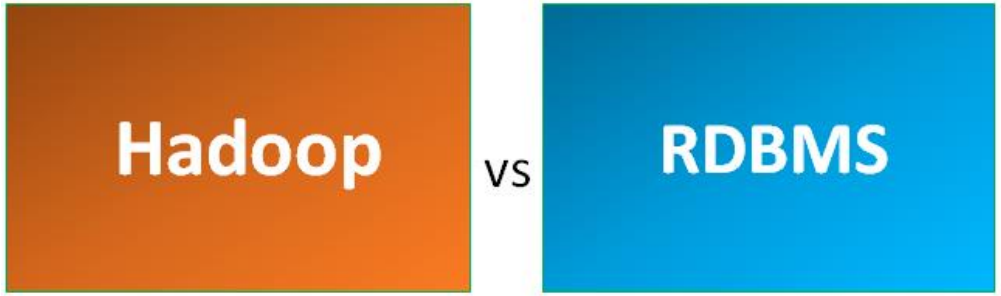
Module Overview

Why Traditional Systems are failing?



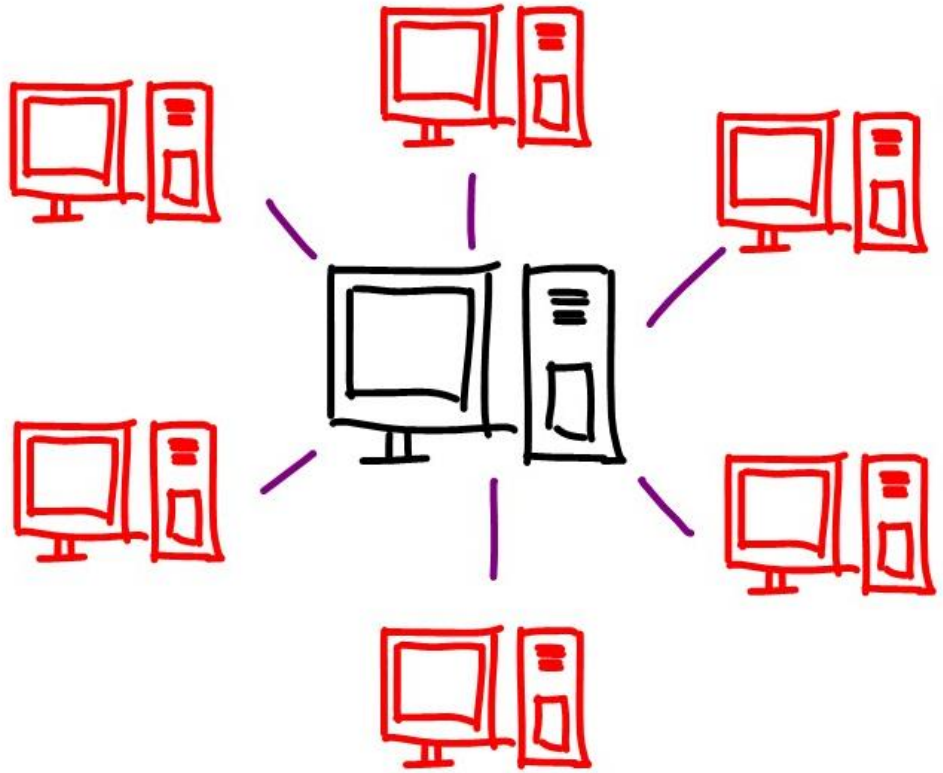
Why Distributed Computing System?

Introducing Hadoop?



Hadoop Ecosystem?

Need of Distributed Computing?

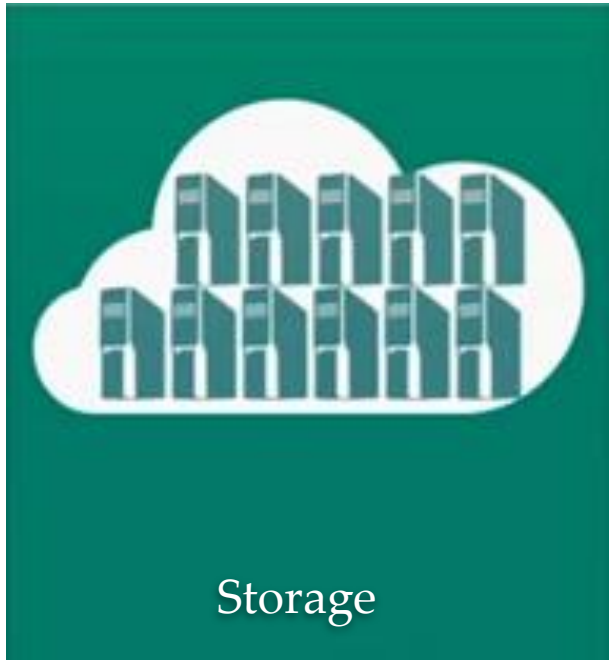


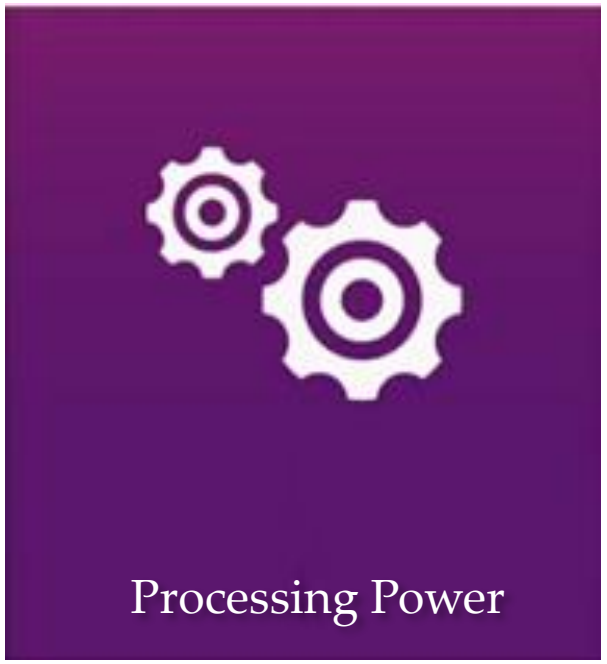
How much data?

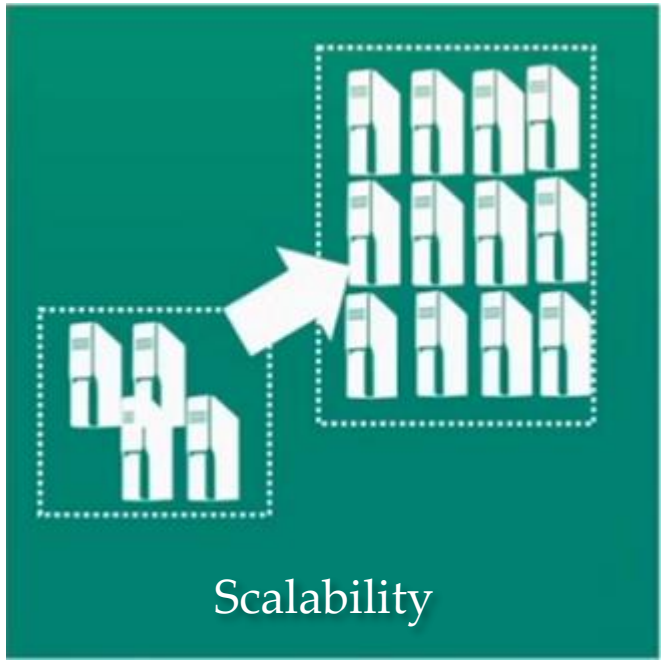
- 2.4 billion monthly active users
- Generate 4 petabytes of data every single day
- 100 million hours of video watch time per hour
- 4 million like every minute

- 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?





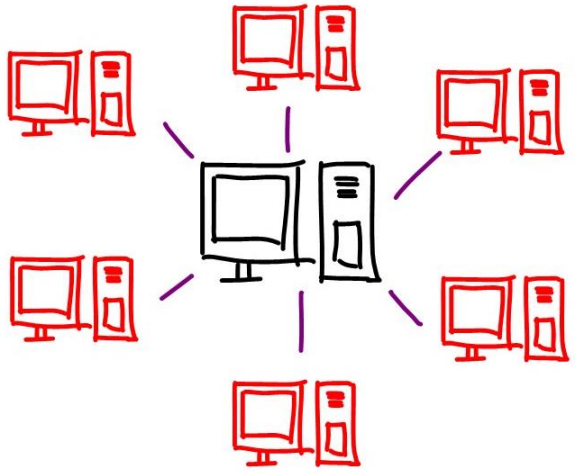


Monolithic system

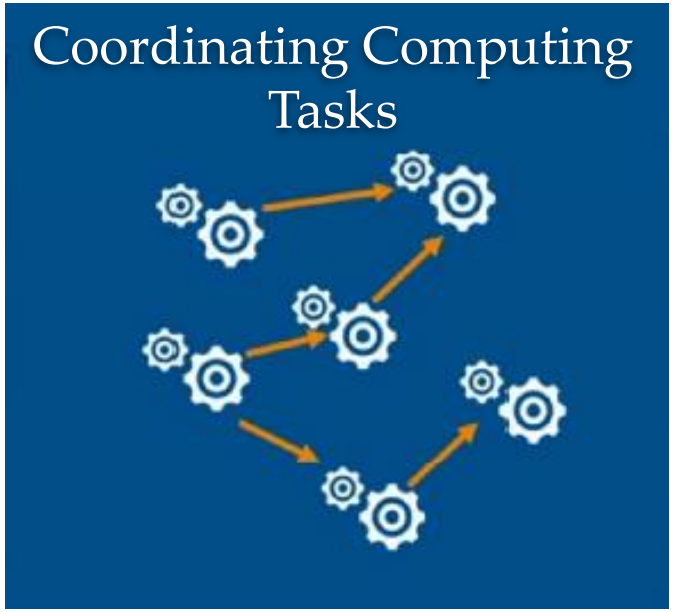
- Single machine
- Single process
- Powerful single server
- Can not scale beyond limit

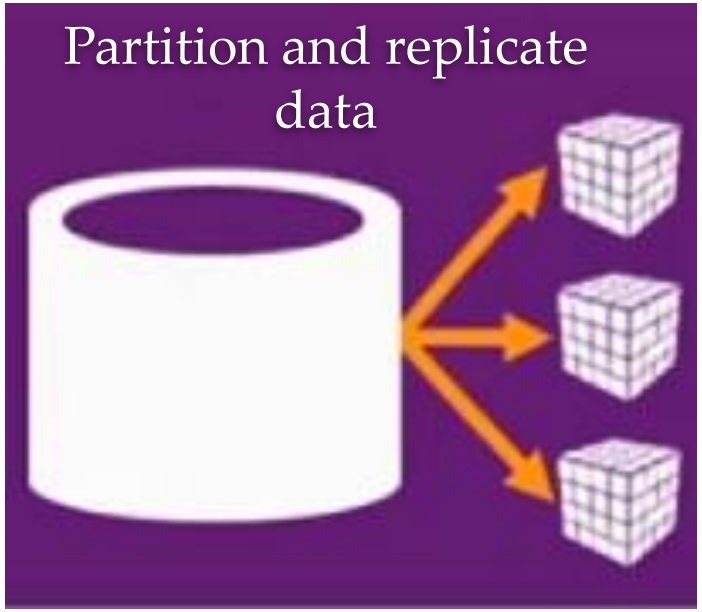
Distributed System

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



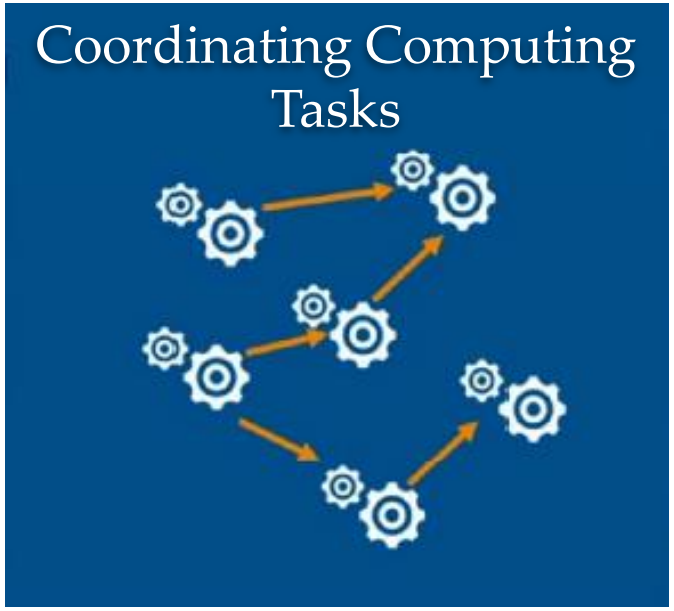
Software requirements to handle Distributed systems?

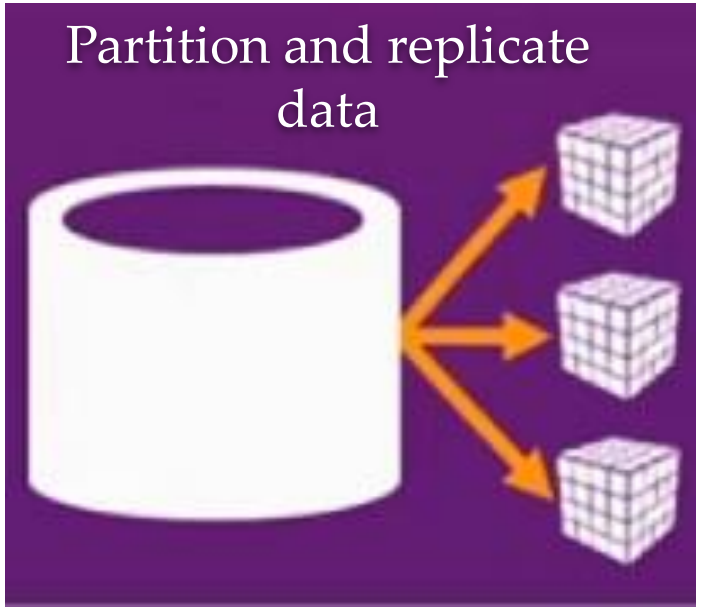




Software purpose is to coordinate and manage all the processes and machines which exist within the system.

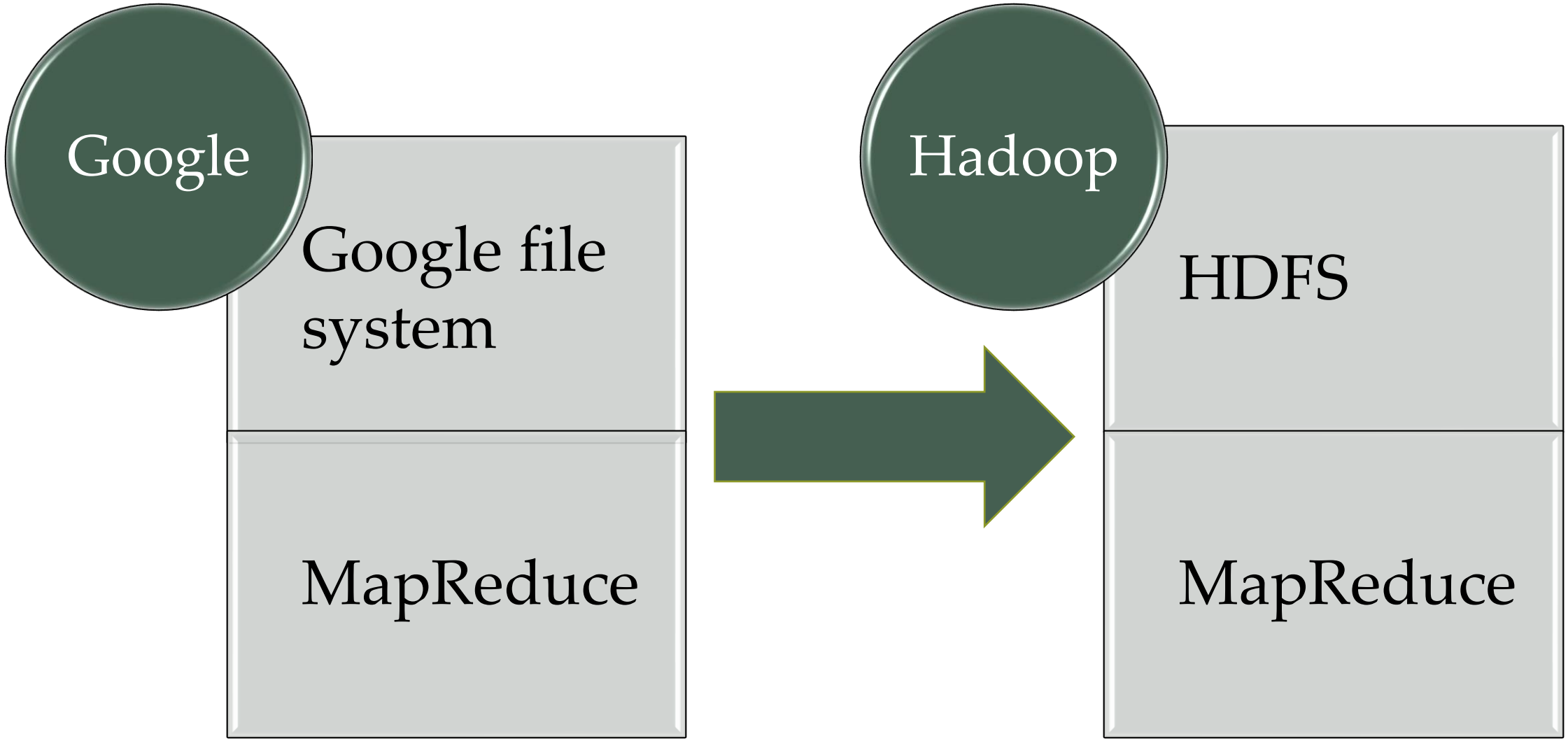
Google Software Challenge?



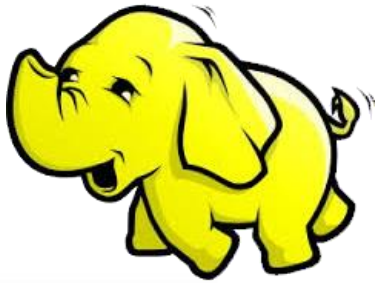


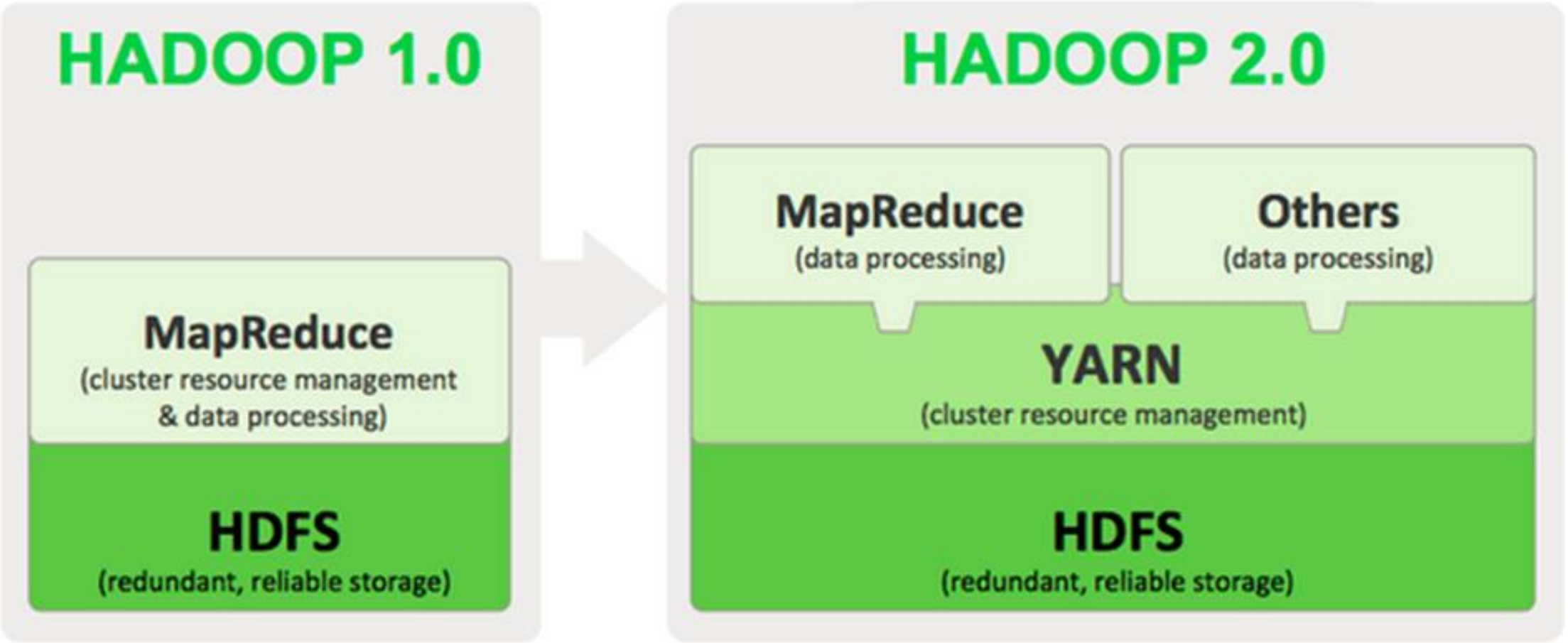
Google software challenge?

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

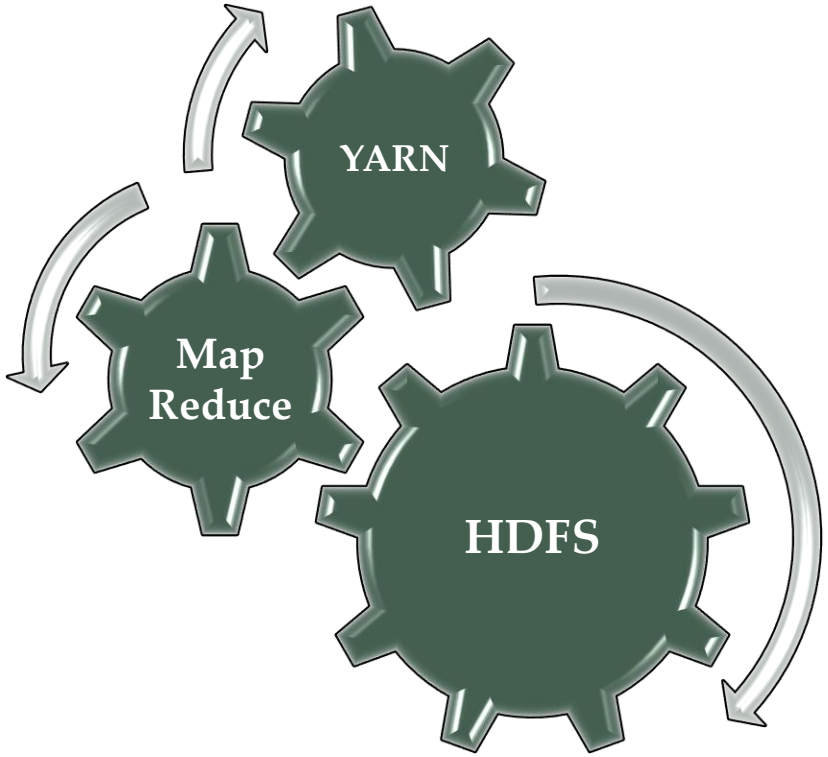

Google
Google file system
MapReduce
Hadoop
HDFS
MapReduce

Hadoop Architecture

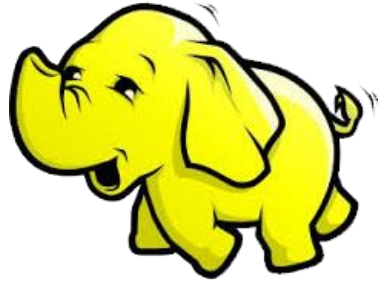


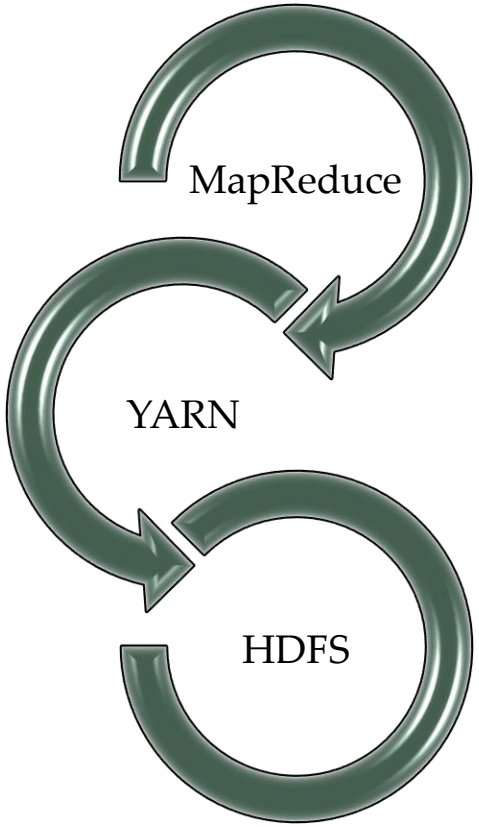


Hadoop Architecture



What happens when you submit a job Hadoop?





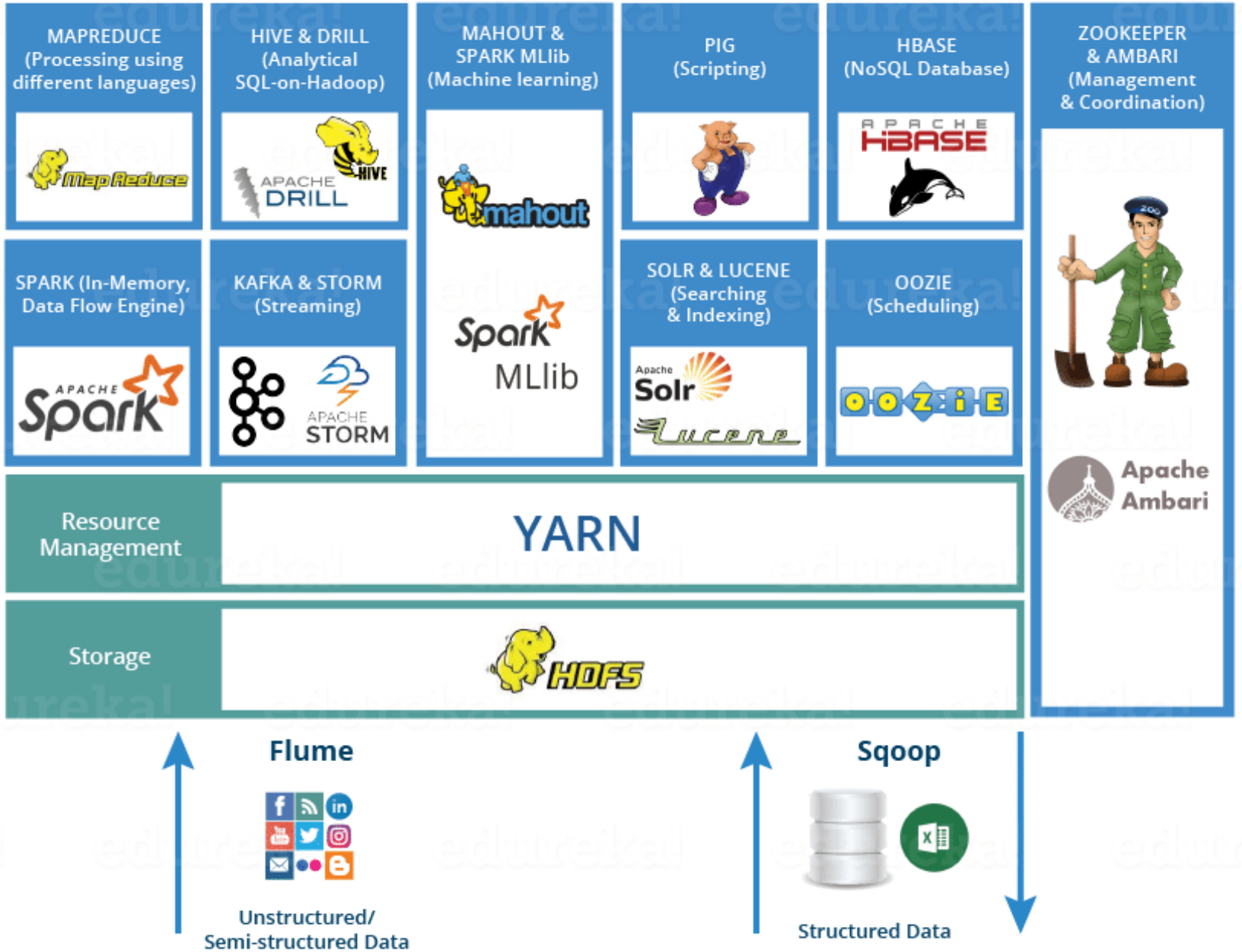
Hadoop vs RDBMS

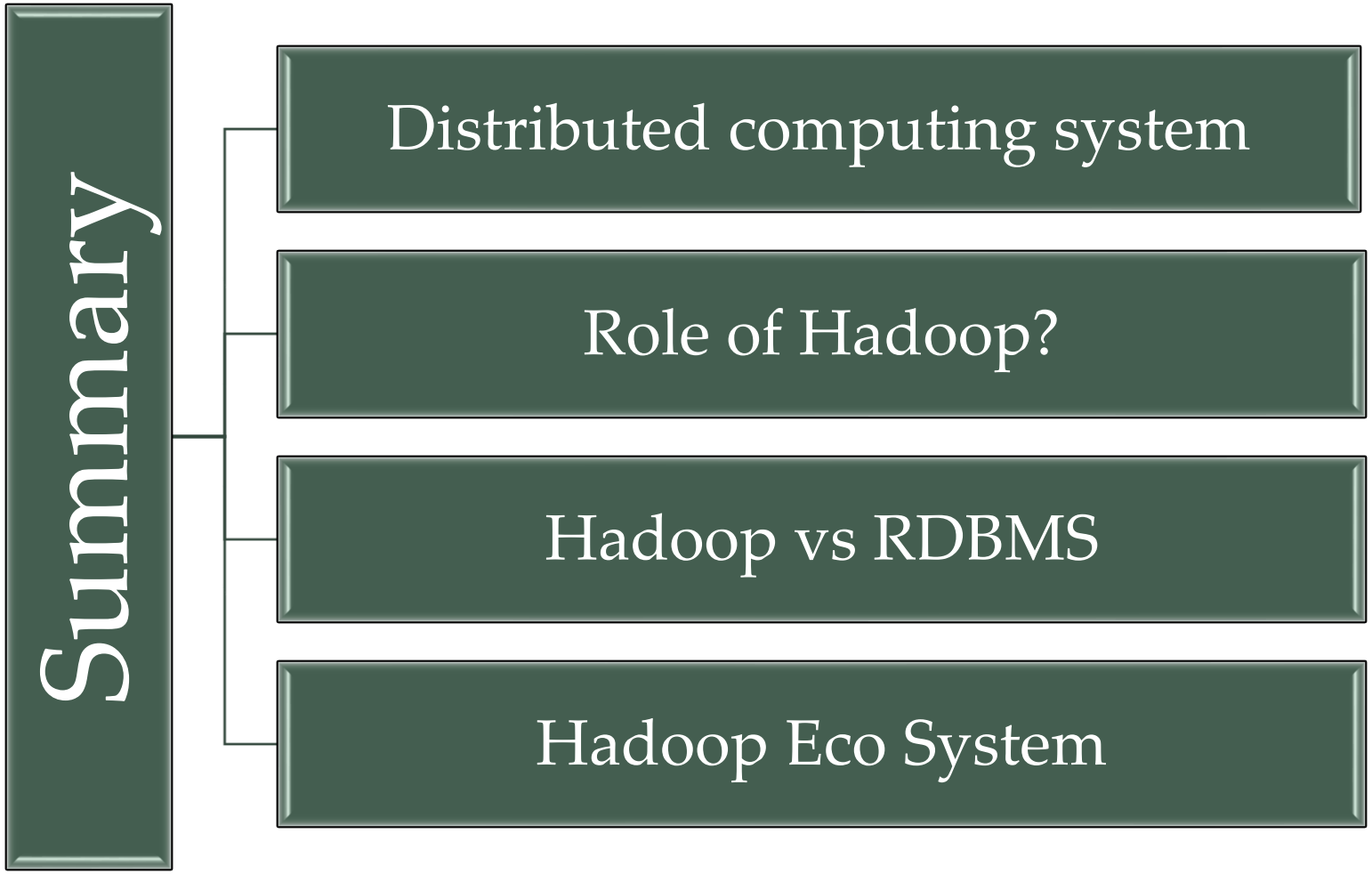
Hadoop

- 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

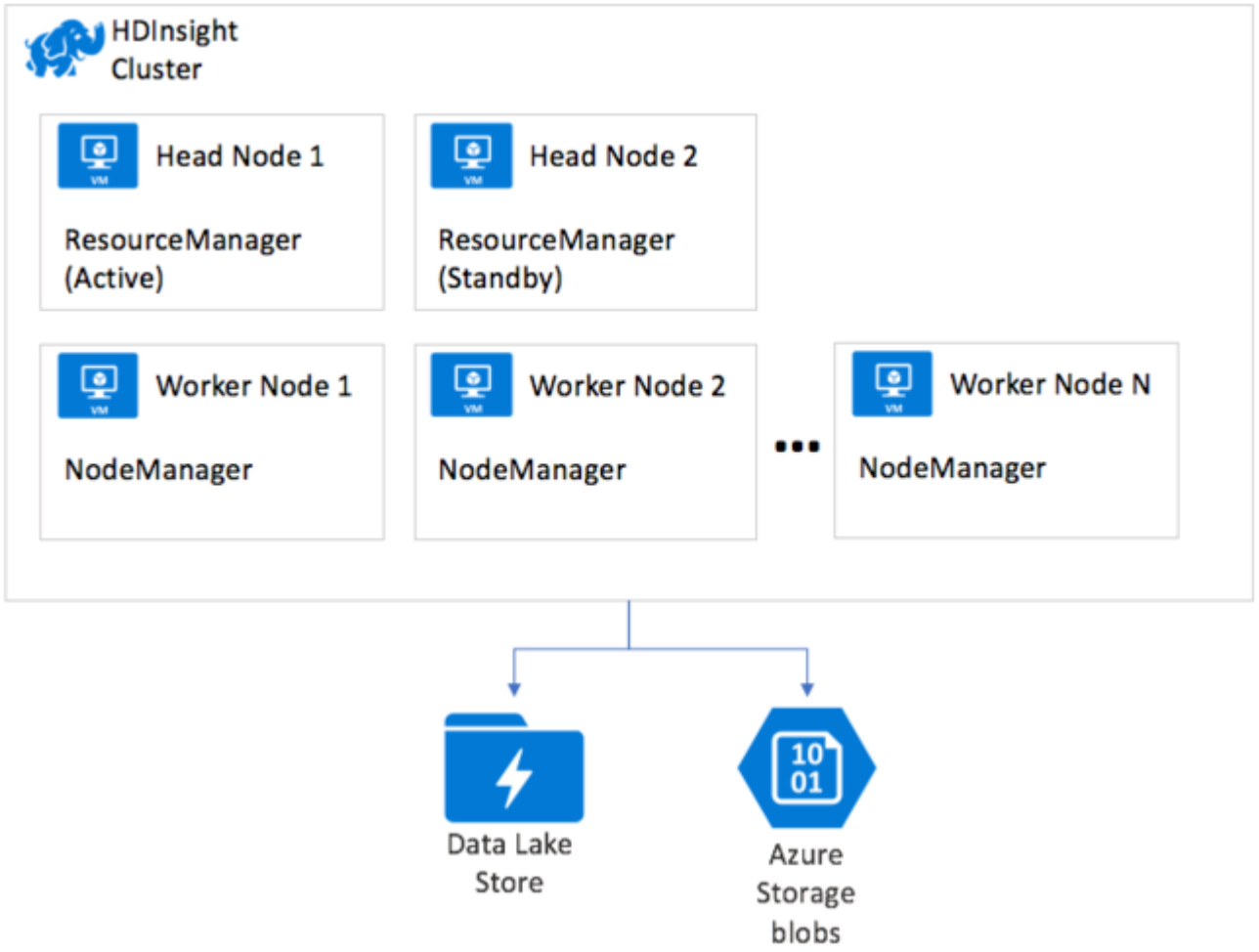

MAPREDUCE
(Processing using different languages)
HIVE & DRILL
(Analytical SQL-on-Hadoop)
MAHOUT & SPARK MLlib
(Machine learning)
PIG
(Scripting)
HBASE
(NoSQL Database)
ZOOKEEPER & AMBARI
(Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
KAFKA & STORM
(Streaming)
SOLR & LUCENE
(Searching & 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

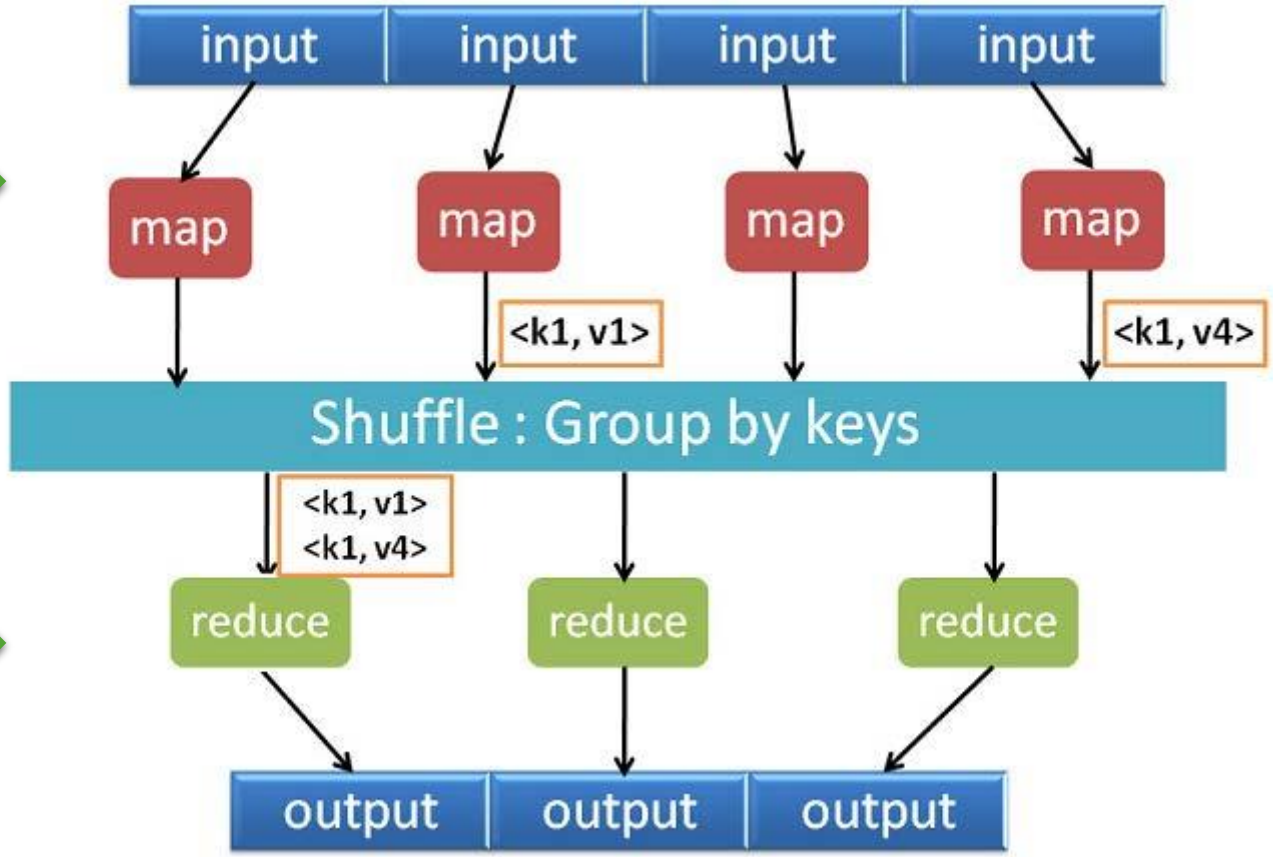
Decoupled Storage



MapReduce operation

Data is chunked redundantly across nodes

Massive Parallelism



Azure HDInsight

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

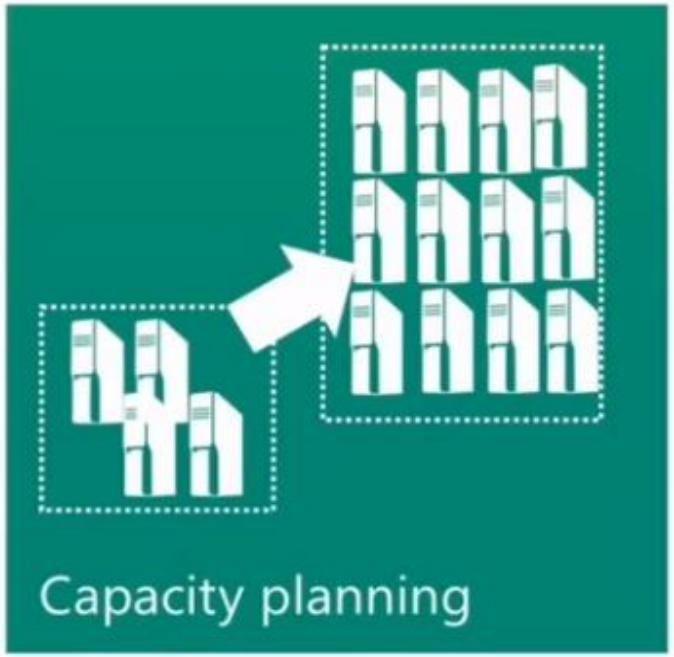
Agenda

- Why Hadoop is Hard?
- Why Hadoop on cloud?
- How HDInsight makes Hadoop easy?
- Important aspects of Hadoop?
- HDInsight Architecture

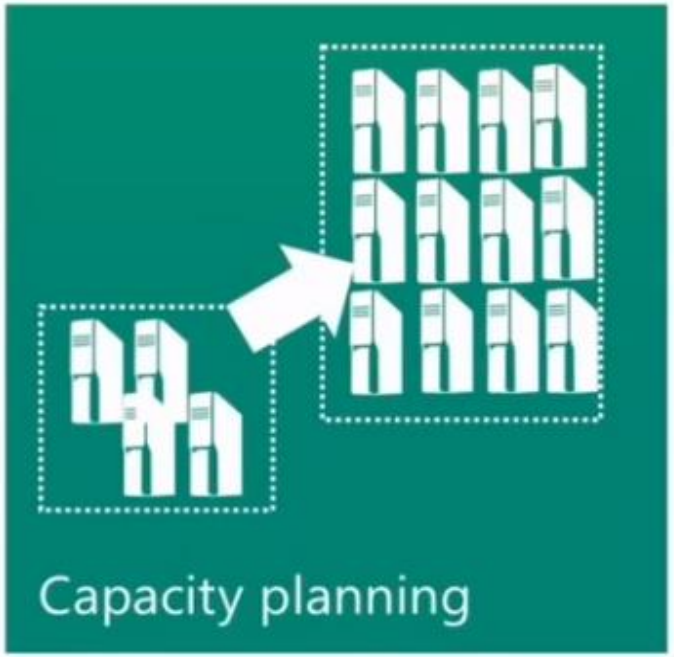
Why Hadoop is Hard?

3 Challenges with Hadoop?

3 Challenges with Hadoop?



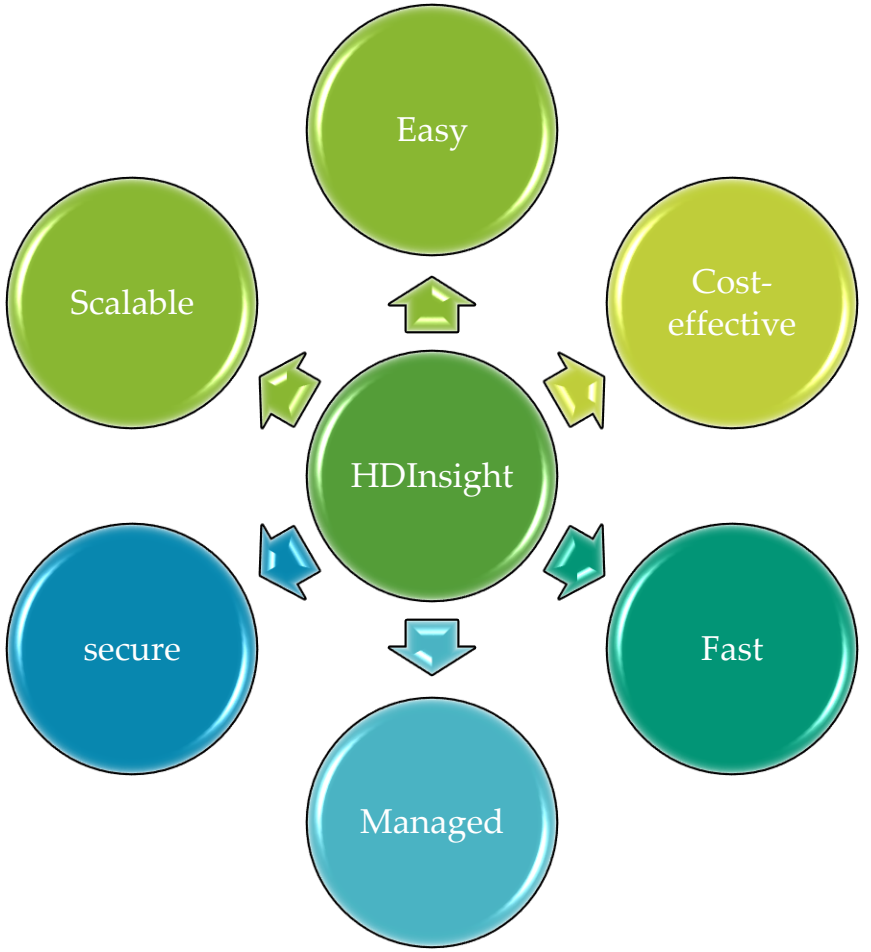
3 Challenges with Hadoop?



HDInsight makes Hadoop easy

What is Azure HDInsight?

HDInsight is a cloud distribution of Hadoop components



HDInsight makes Hadoop easy

HDInsight makes Hadoop easy

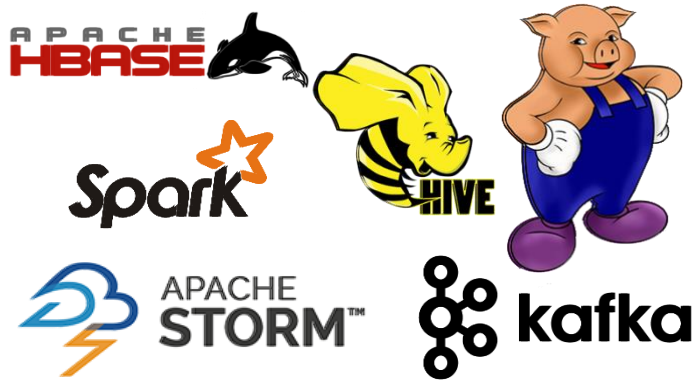
HDInsight makes Hadoop easy

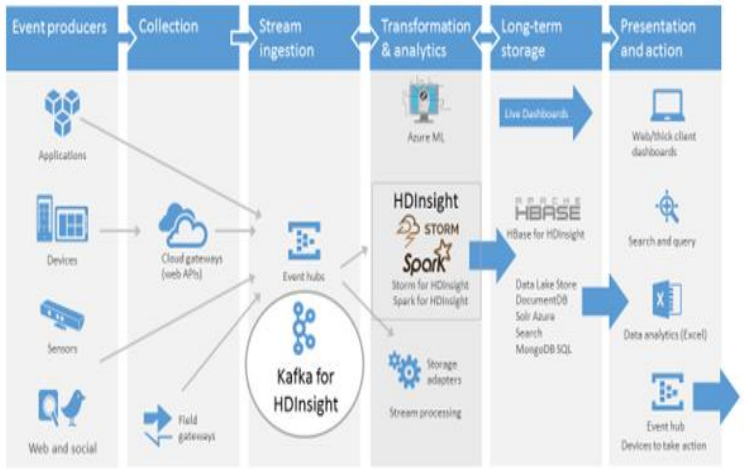
HDInsight makes Hadoop easy

Important aspects of HDInsight

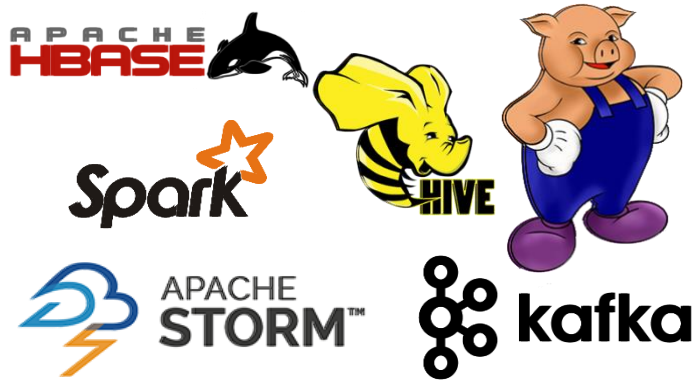
Important aspects of HDInsight

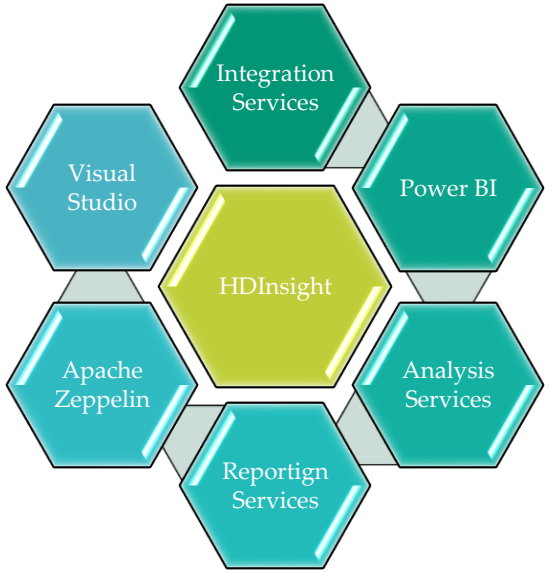
Important aspects of HDInsight

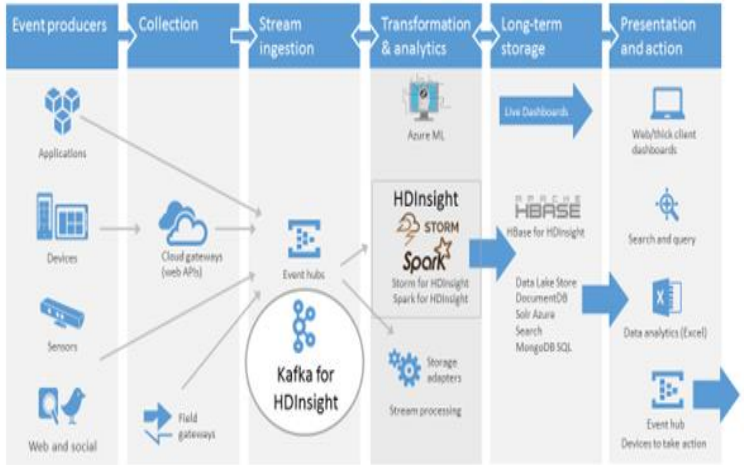




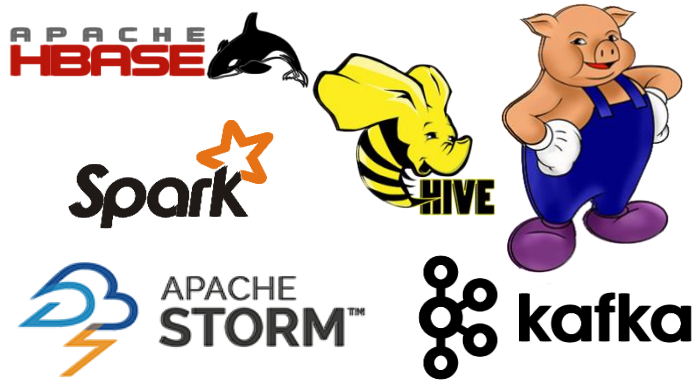
Important aspects of HDInsight

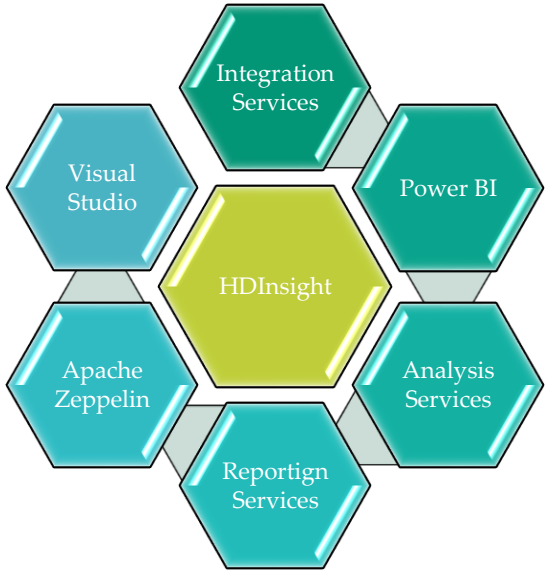


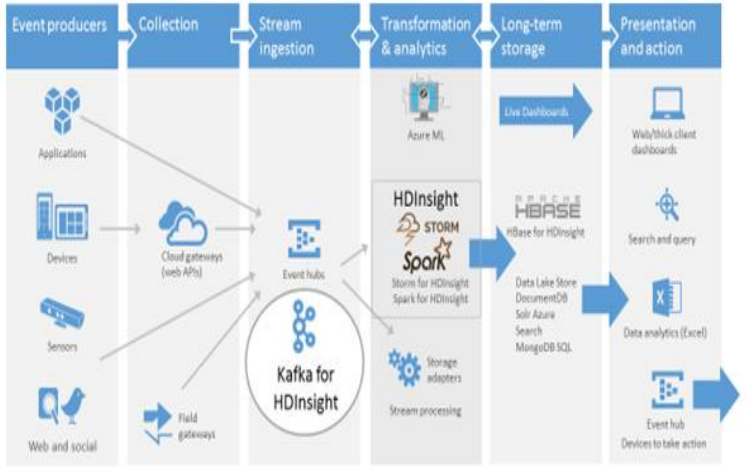


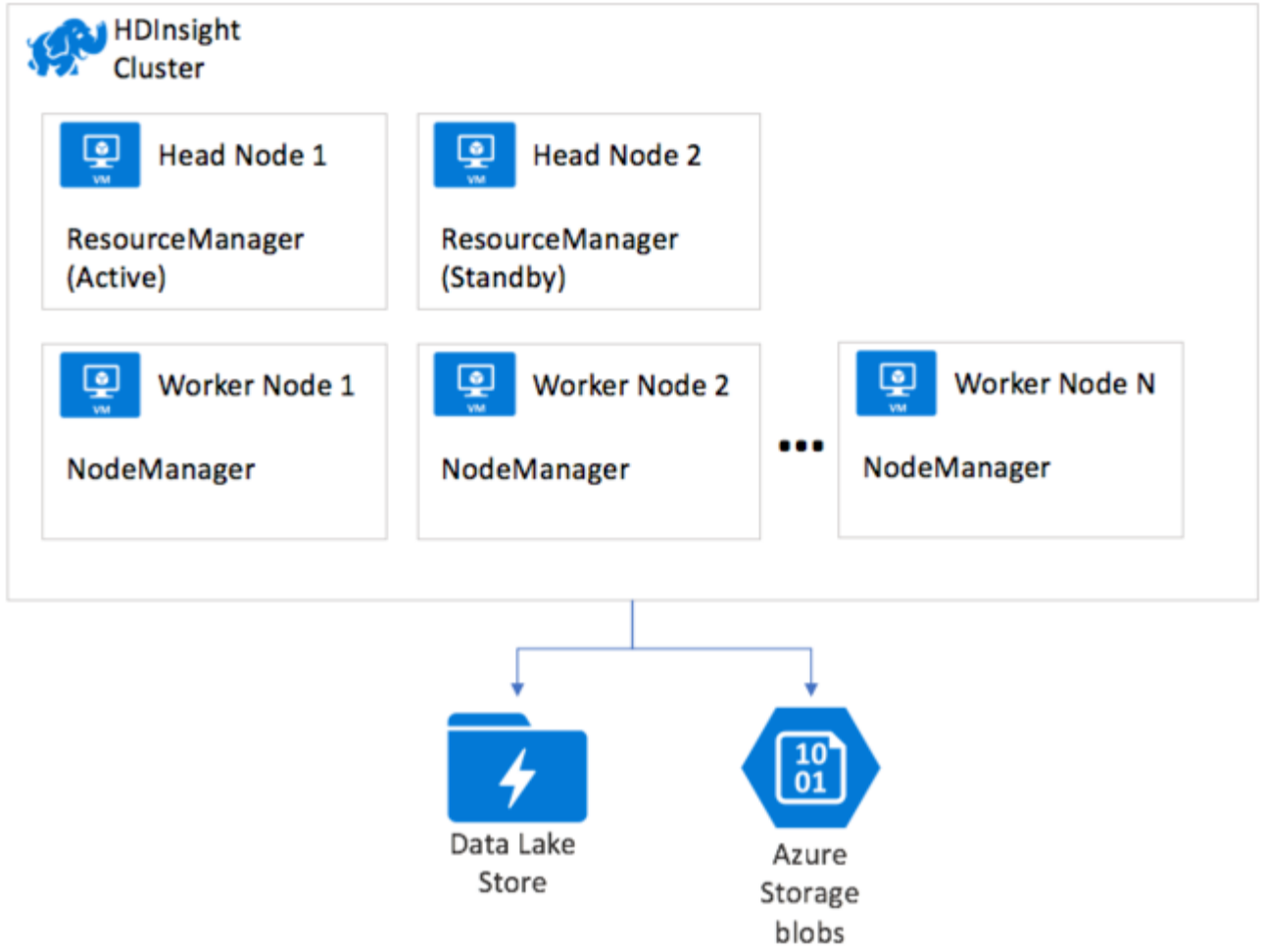


Important aspects of HDInsight



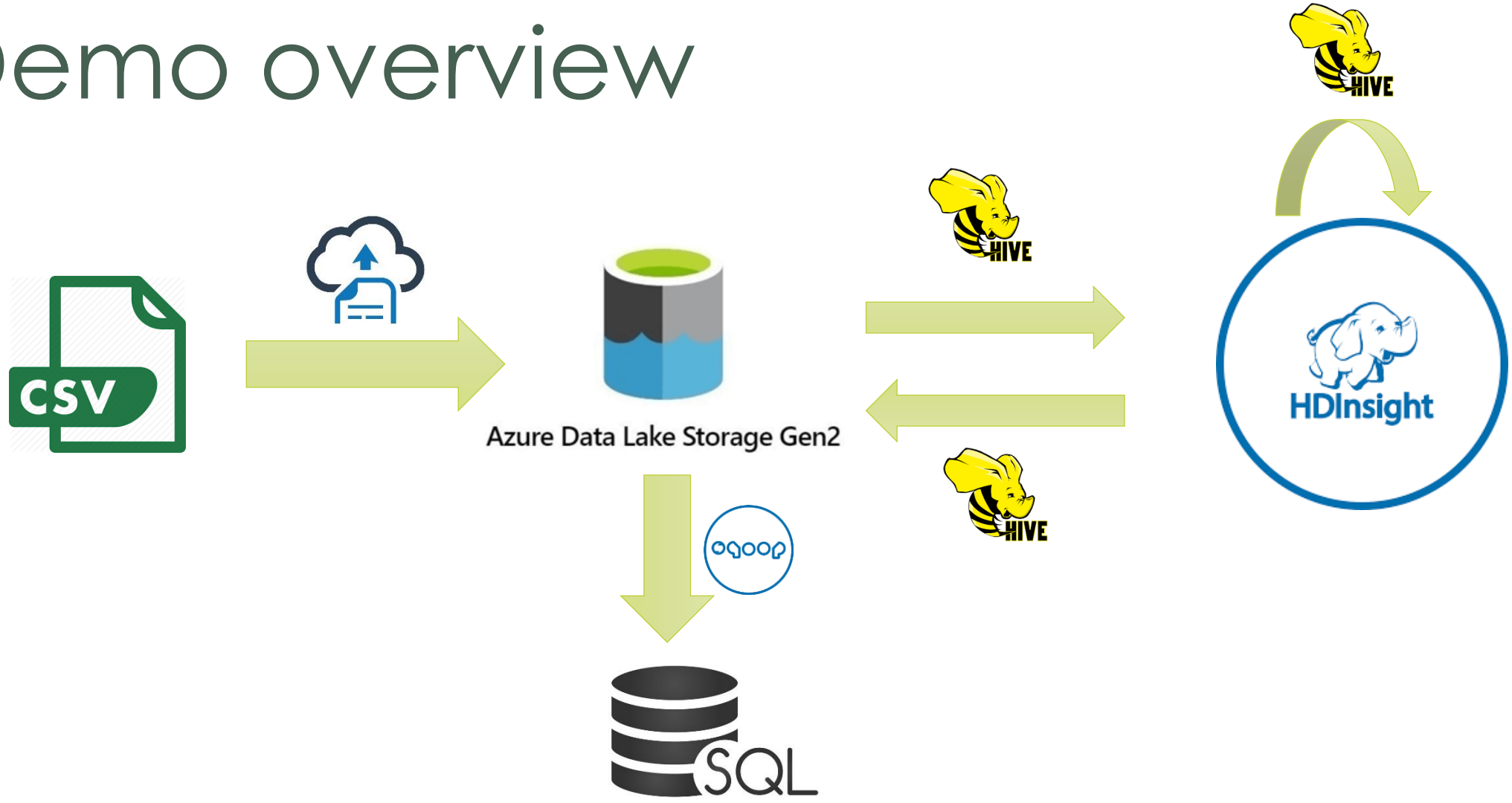





HDInsight
Cluster
Head Node 1
ResourceManager
(Active)
Head Node 2
ResourceManager
(Standby)
Worker Node 1
NodeManager
Worker Node 2
NodeManager
Worker Node N
NodeManager
Data Lake
Store
Azure
Storage
blobs

Demo - HDInsight

Demo overview



Ambari

Ambari

"Ambari is a Hadoop management platform responsible for cluster administration, monitoring and configuration"

Flight traffic control

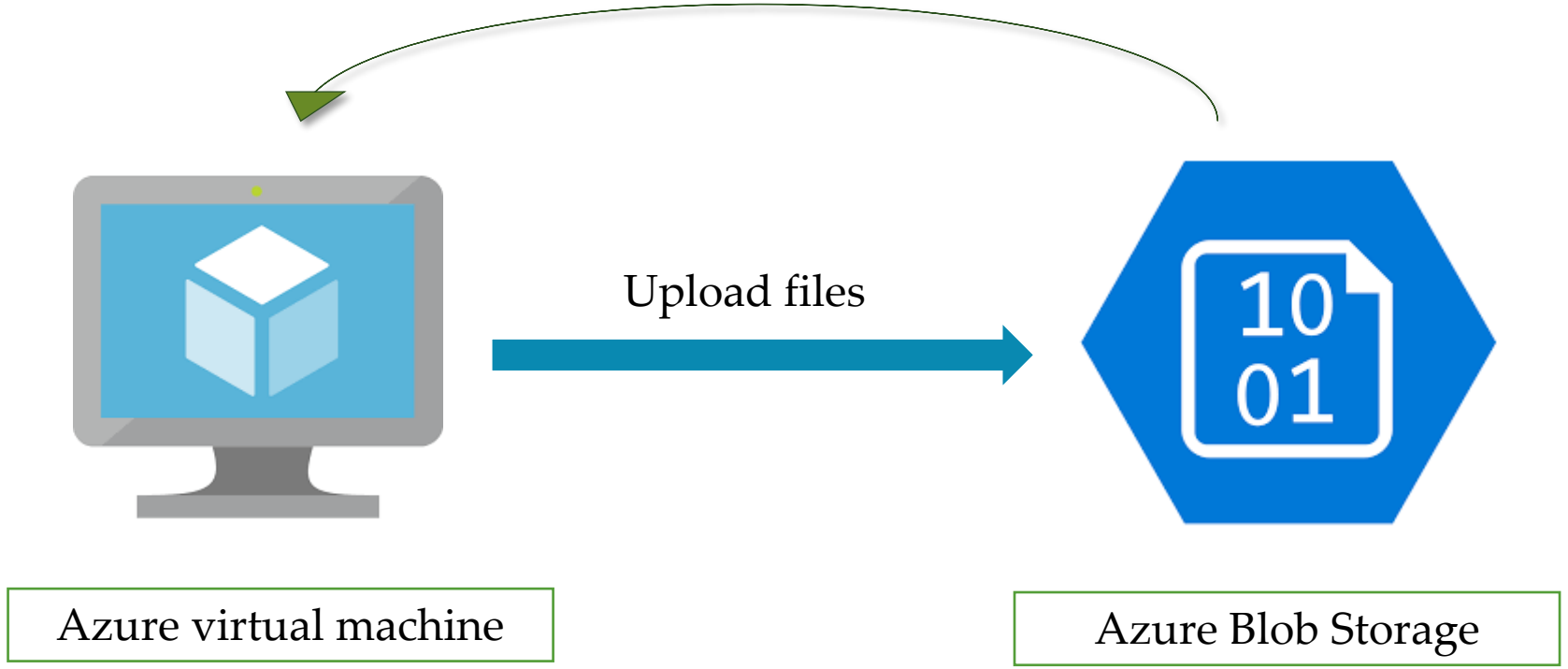
- Visualized reports
- System monitoring
- Change configuration
- Manage authentication

Managed Identity

Managed Identity

Managed identities are
used by Azure services to authenticate to other Azure services
that support Azure AD authentication.

Managed Identity



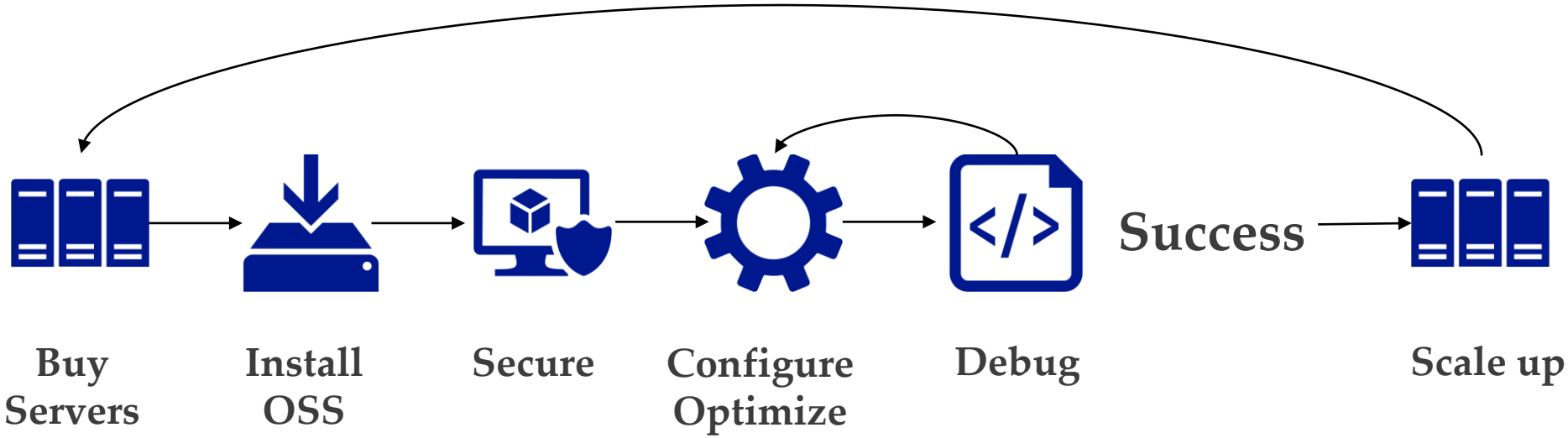
Managed Identity – 2 steps

Two types of Managed Identity

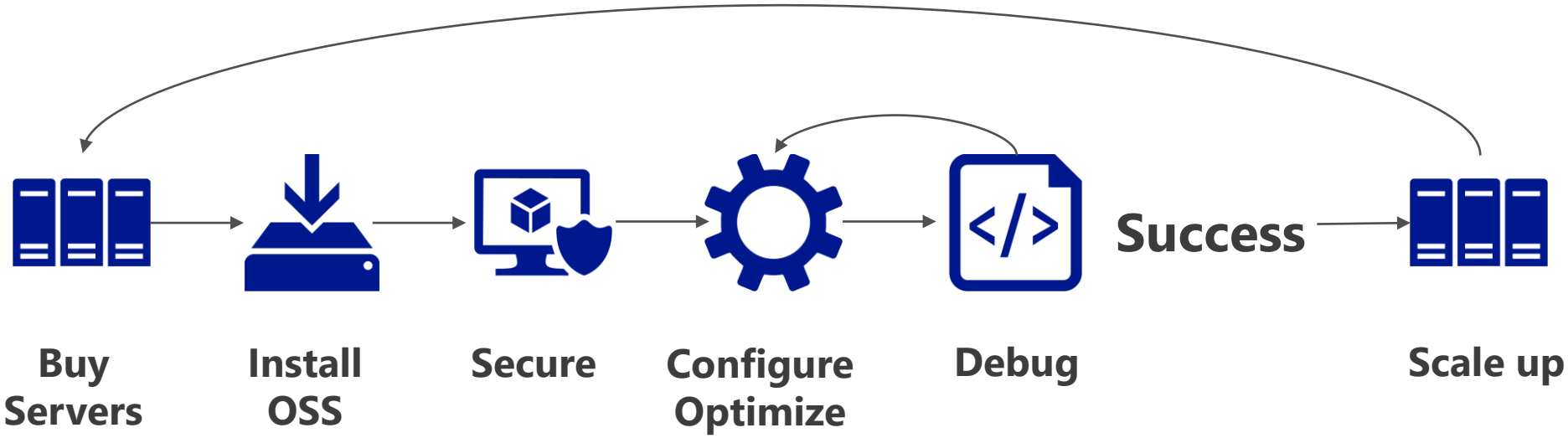
System-assigned	• Enable directly on Azure service instance • Lifecycle is tied to service instance
User-assigned	• Created as a stand alone Azure resource • Lifecycle managed separately

Summary

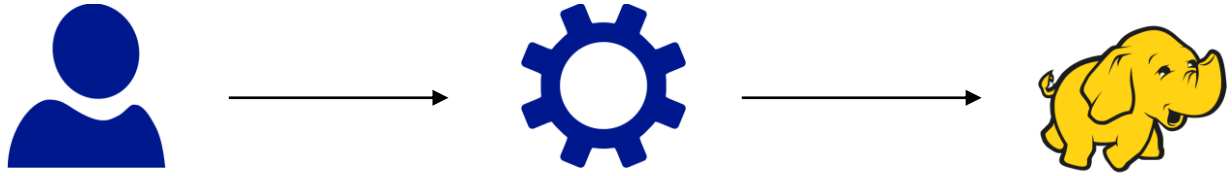
Big Data Was Hard?



Big data is hard



HDInsight makes it easy



Browse to Azure Portal

Provide Cluster details

HDInsight Cluster

- ✓ **100% open source**
- ✓ **Optimized**
- ✓ **Highly available**
- ✓ **Secure**
- ✓ **Scalable**
- ✓ **Dedicated**
- ✓ **Managed**
- ✓ **Certified ISVs**
- ✓ **Customizable**

Rich Developer Ecosystem

Plugins for HDI available for most popular IDEs for agile development and debugging

Rich support for powerful notebooks used by data scientists

Develop in C#, deploy on Linux in Java via HDI developed SCP.Net technology

Open Source for the Enterprise

Managed Open Source Analytics for the cloud with a 99.9% SLA.

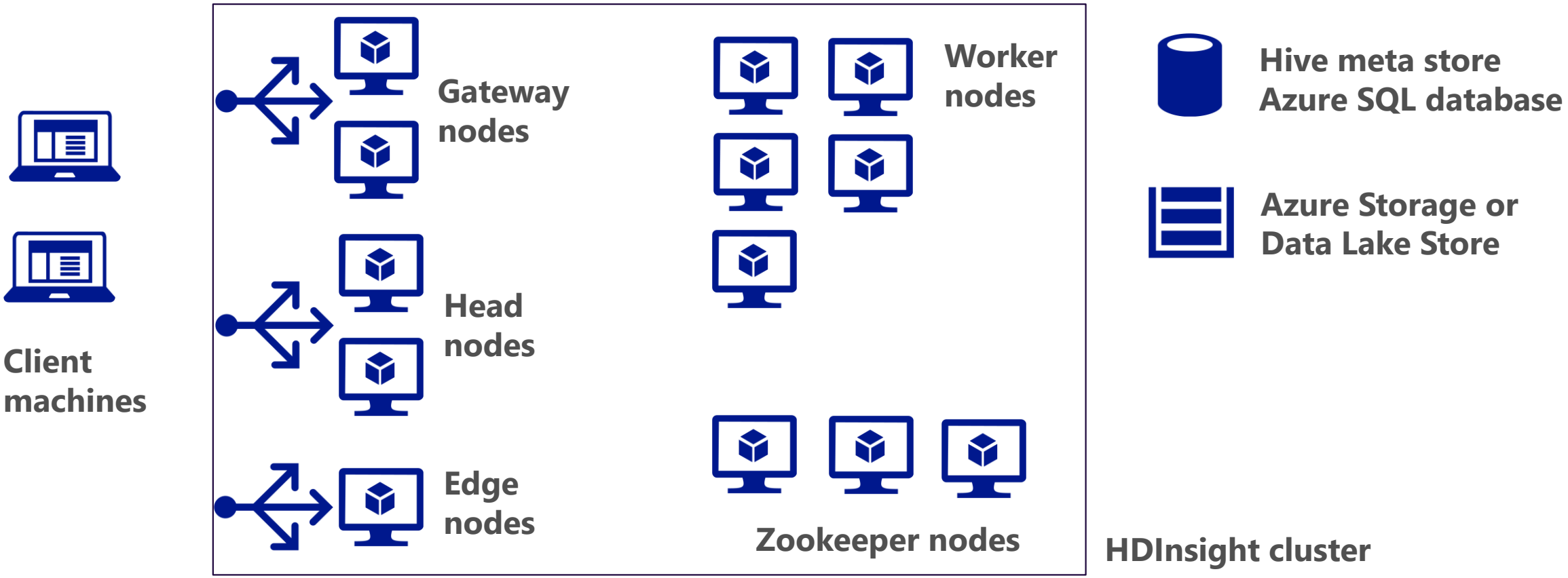
100% Open Source

Clusters up and running in minutes

63% lower TCO than deploy your own Hadoop on-premises

Separation of compute and store allows you to scale clusters to exponentially reduce costs

HDInsight architecture



Why Hadoop in the cloud?

- Cloud distribution of Hadoop components
- Easy, fast, and cost-effective to process massive amounts of data
- **Fully managed service based on node size and type**
- **Enterprise grade security**
-
- **Spark** – In-memory processing and interactive queries
- **Strom** – Real-time event processing

HDInsight cluster types

- **Hadoop** – Batch query and analysis of HDFS stored data
- **HBase** – Processing for large schema less NoSQL data
- **Interactive Query** – In-memory caching for fast Hive queries
- **Kafka** – Distributed streaming data platform
- **ML Services** – Predictive modeling and machine learning
- **Spark** – In-memory processing and interactive queries
- **Strom** – Real-time event processing