The Hadoop Ecosystem

HORTONWORKS UNIVERSITY

Lesson Objectives

After completing this lesson, students should be able to:

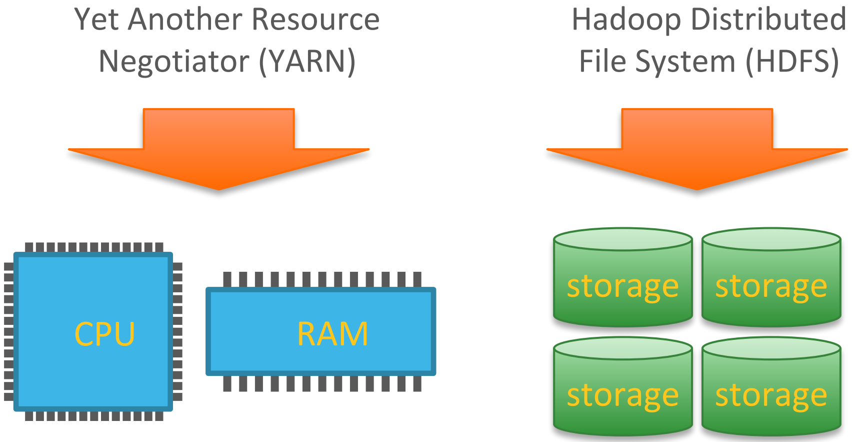
- Describe the Hadoop ecosystem frameworks across the following five architectural categories:
  - Data Management
  - Data Access
  - Data Governance & Integration
  - Security
  - Operations
- Deploy Hadoop into a datacenter – Connected Data Platforms
  - Hadoop cluster node types
  - Integrating with existing data applications

Hadoop Ecosystem Frameworks

Hadoop in the Datacenter

HORTONWORKS UNIVERSITY

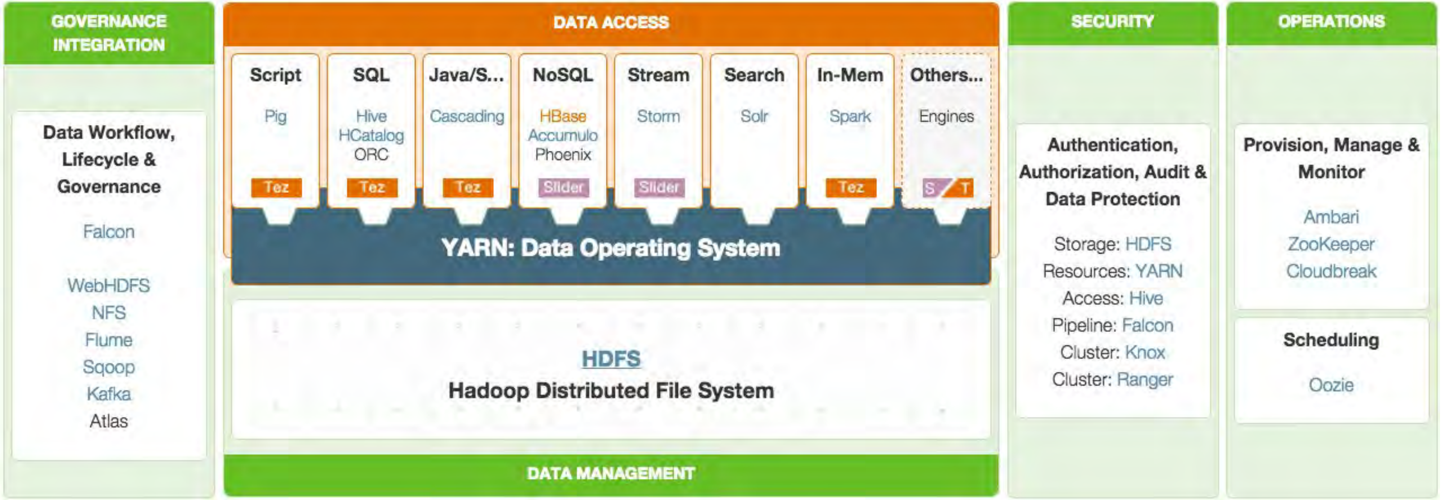
Hadoop Core = Storage + Compute


Yet Another Resource Negotiator (YARN)
Hadoop Distributed File System (HDFS)
CPU
RAM
storage
storage
storage
storage

The Hadoop Ecosystem

Apache Ambari
http://incubator.apache.org/ambari
Spark
TEZ
sqoop
APACHE FALCON
Apache orc™
APACHE KNOX
HIVE
Hadoop
OOZIE
Apache Solr
APACHE HBASE
accumulo
STORM
kafka
ZOO
FLUME

Hortonworks Hadoop Distribution



Data Management Frameworks

Framework	Description
Hadoop Distributed File System (HDFS)	A Java-based, distributed file system that provides scalable, reliable, high-throughput access to application data stored across commodity servers
Yet Another Resource Negotiator (YARN)	A framework for cluster resource management and job scheduling

Operations Frameworks

Framework	Description
Ambari	A Web-based framework for provisioning, managing, and monitoring Hadoop clusters
ZooKeeper	A high-performance coordination service for distributed applications
Cloudbreak	A tool for provisioning and managing Hadoop clusters in the cloud
Oozie	A server-based workflow engine used to execute Hadoop jobs

Data Access Frameworks

Framework	Description
Pig	A high-level platform for extracting, transforming, or analyzing large datasets
Hive	A data warehouse infrastructure that supports ad hoc SQL queries
HCatalog	A table information, schema, and metadata management layer supporting Hive, Pig, MapReduce, and Tez processing
Cascading	An application development framework for building data applications, abstracting the details of complex MapReduce programming
HBase	A scalable, distributed NoSQL database that supports structured data storage for large tables
Phoenix	A client-side SQL layer over HBase that provides low-latency access to HBase data
Accumulo	A low-latency, large table data storage and retrieval system with cell-level security
Storm	A distributed computation system for processing continuous streams of real-time data
Solr	A distributed search platform capable of indexing petabytes of data
Spark	A fast, general purpose processing engine use to build and run sophisticated SQL, streaming, machine learning, or graphics applications

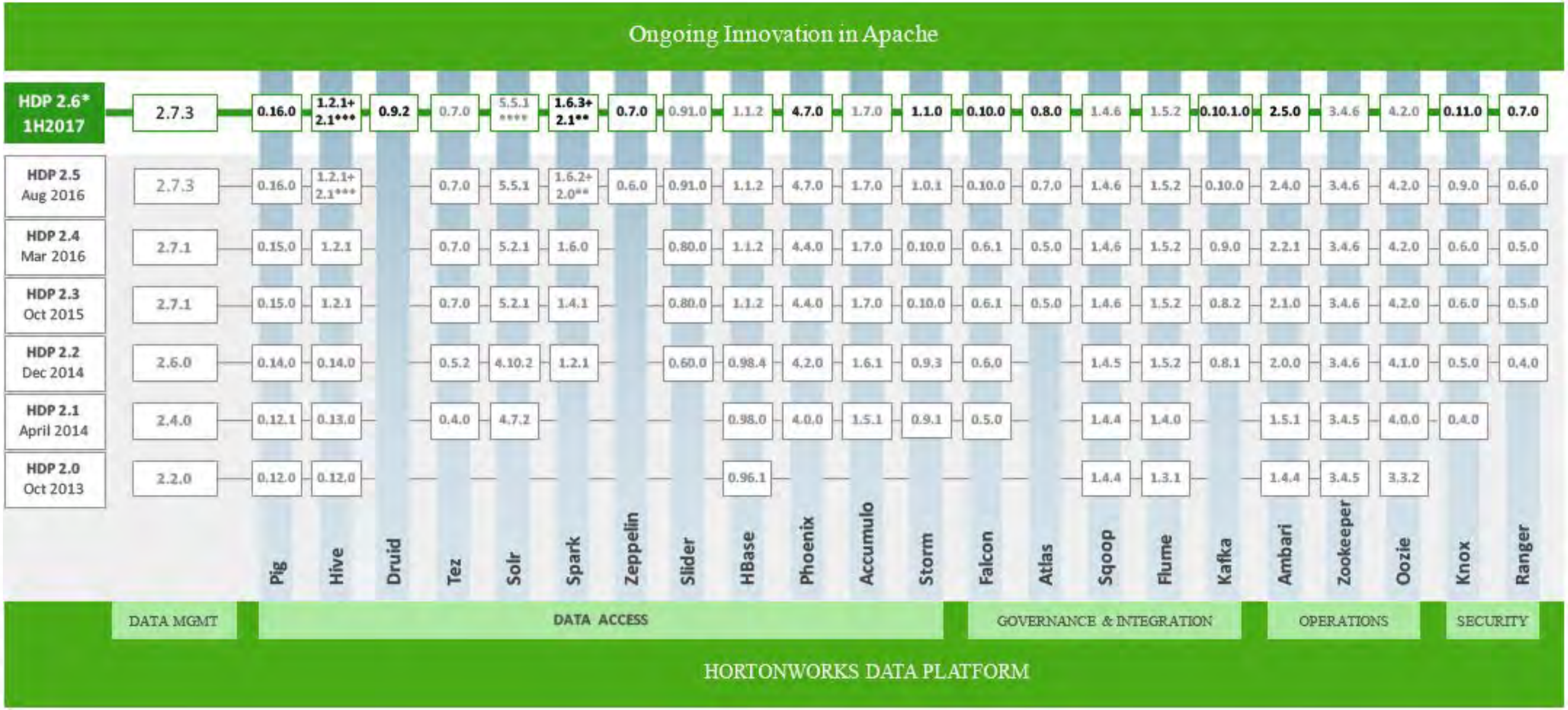
Governance and Integration Frameworks

Framework	Description
Falcon	A data governance tool providing workflow orchestration, data lifecycle management, and data replication services.
WebHDFS	A REST API that uses the standard HTTP verbs to access, operate, and manage HDFS
HDFS NFS Gateway	A gateway that enables access to HDFS as an NFS mounted file system
Flume	A distributed, reliable, and highly-available service that efficiently collects, aggregates, and moves streaming data
Sqoop	A set of tools for importing and exporting data between Hadoop and RDBM systems
Kafka	A fast, scalable, durable, and fault-tolerant publish-subscribe messaging system
Atlas	A scalable and extensible set of core governance services enabling enterprises to meet compliance and data integration requirements

Security Frameworks

Framework	Description
HDFS	A storage management service providing file and directory permissions, even more granular file and directory access control lists, and transparent data encryption
YARN	A resource management service with access control lists controlling access to compute resources and YARN administrative functions
Hive	A data warehouse infrastructure service providing granular access controls to table columns and rows
Falcon	A data governance tool providing access control lists that limit who may submit Hadoop jobs
Knox	A gateway providing perimeter security to a Hadoop cluster
Ranger	A centralized security framework offering fine-grained policy controls for HDFS, Hive, HBase, Knox, Storm, Kafka, and Solr

Ecosystem Component Versions



Ongoing Innovation in Apache

		Pig	Hive	Druid	Tez	Solr	Spark	Zeppelin	Slider	HBase	Phoenix	Accumulo	Storm	Falcon	Atlas	Sqoop	Flume	Kafka	Ambari	Zookeeper	Oozie	Knox	Ranger
HDP 2.6* 1H2017	2.7.3	0.16.0	1.2.1+ 2.1***	0.9.2	0.7.0	5.5.1 ****	1.6.3+ 2.1**	0.7.0	0.91.0	1.1.2	4.7.0	1.7.0	1.1.0	0.10.0	0.8.0	1.4.6	1.5.2	0.10.1.0	2.5.0	3.4.6	4.2.0	0.11.0	0.7.0
HDP 2.5 Aug 2016	2.7.3	0.16.0	1.2.1+ 2.1***		0.7.0	5.5.1	1.6.2+ 2.0**	0.6.0	0.91.0	1.1.2	4.7.0	1.7.0	1.0.1	0.10.0	0.7.0	1.4.6	1.5.2	0.10.0	2.4.0	3.4.6	4.2.0	0.9.0	0.6.0
HDP 2.4 Mar 2016	2.7.1	0.15.0	1.2.1		0.7.0	5.2.1	1.6.0		0.80.0	1.1.2	4.4.0	1.7.0	0.10.0	0.6.1	0.5.0	1.4.6	1.5.2	0.9.0	2.2.1	3.4.6	4.2.0	0.6.0	0.5.0
HDP 2.3 Oct 2015	2.7.1	0.15.0	1.2.1		0.7.0	5.2.1	1.4.1		0.80.0	1.1.2	4.4.0	1.7.0	0.10.0	0.6.1	0.5.0	1.4.6	1.5.2	0.8.2	2.1.0	3.4.6	4.2.0	0.6.0	0.5.0
HDP 2.2 Dec 2014	2.6.0	0.14.0	0.14.0		0.5.2	4.10.2	1.2.1		0.60.0	0.98.4	4.2.0	1.6.1	0.9.3	0.6.0		1.4.5	1.5.2	0.8.1	2.0.0	3.4.6	4.1.0	0.5.0	0.4.0
HDP 2.1 April 2014	2.4.0	0.12.1	0.13.0		0.4.0	4.7.2				0.98.0	4.0.0	1.5.1	0.9.1	0.5.0		1.4.4	1.4.0		1.5.1	3.4.5	4.0.0	0.4.0	
HDP 2.0 Oct 2013	2.2.0	0.12.0	0.12.0							0.96.1						1.4.4	1.3.1		1.4.4	3.4.5	3.3.2		

DATA MGMT

DATA ACCESS

GOVERNANCE & INTEGRATION

OPERATIONS

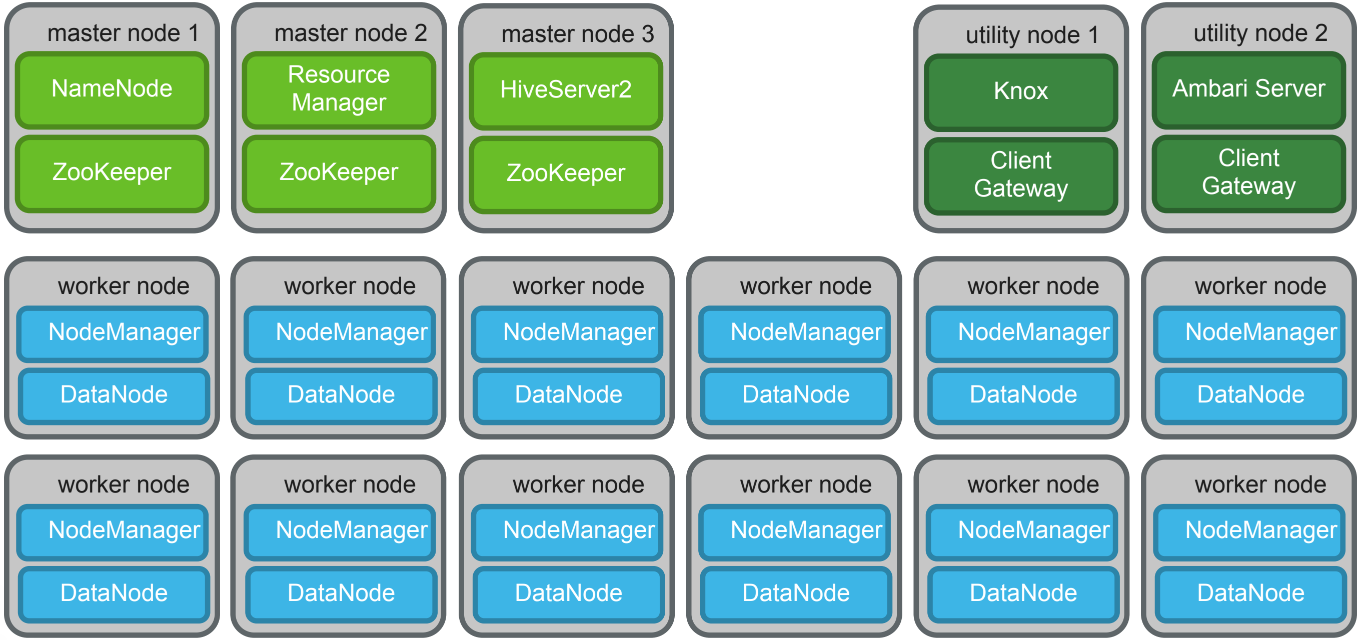
SECURITY

HORTONWORKS DATA PLATFORM

Hadoop Ecosystem Frameworks

Hadoop in the Datacenter

Distinct Masters and Scale-Out Workers



master node 1
NameNode
ZooKeeper

master node 2
Resource Manager
ZooKeeper

master node 3
HiveServer2
ZooKeeper

utility node 1
Knox
Client Gateway

utility node 2
Ambari Server
Client Gateway

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

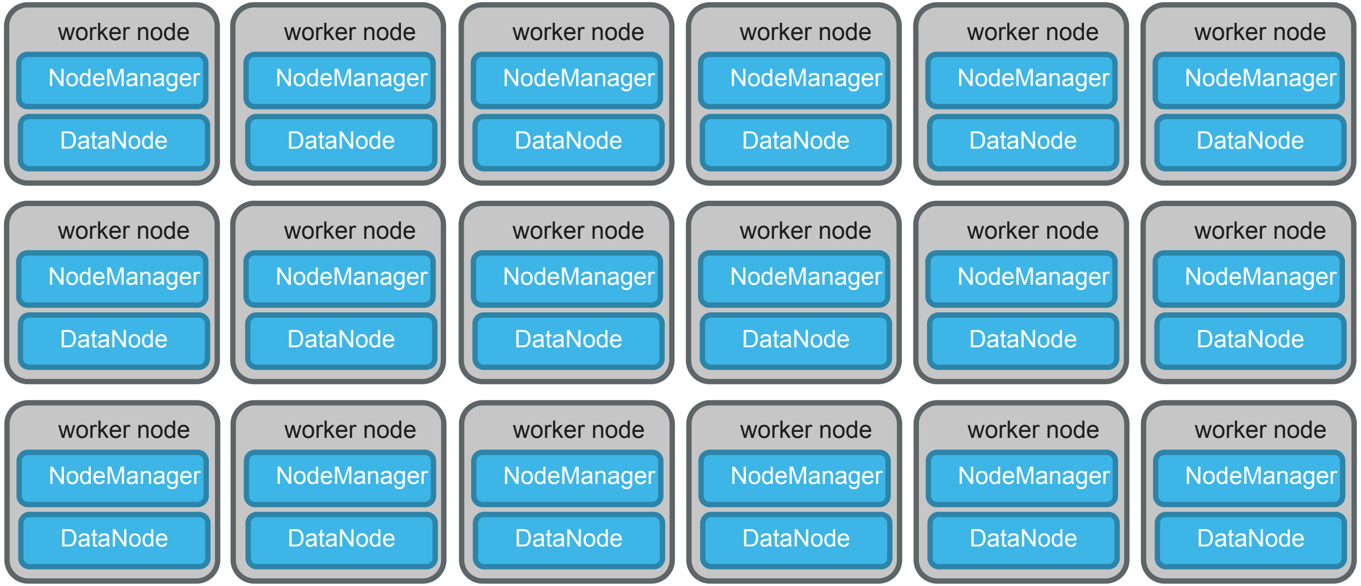
worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

worker node
NodeManager
DataNode

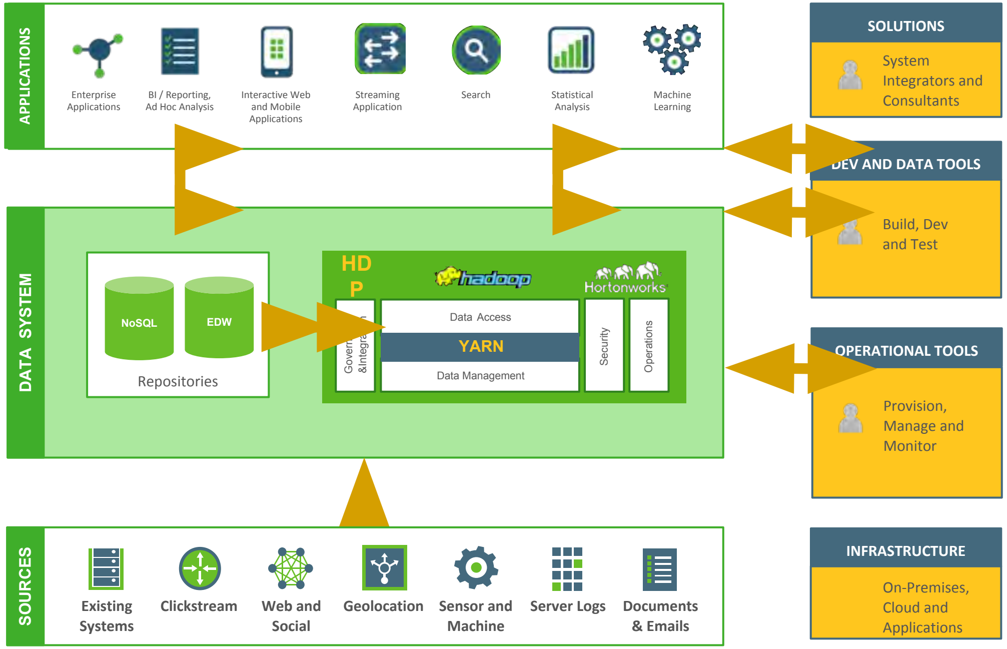
worker node
NodeManager
DataNode

Worker Nodes can Scale into the Thousands

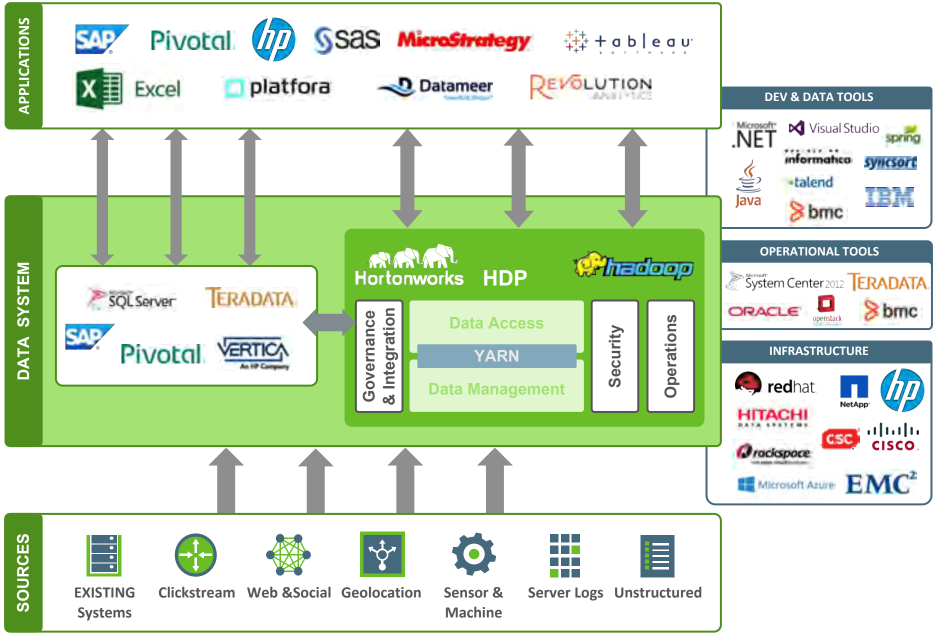


worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode
worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode
worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode	worker node NodeManager DataNode

Connected Data Platforms



Hadoop as a +1 Architecture



APPLICATIONS

DATA SYSTEM

SOURCES

DEV & DATA TOOLS

OPERATIONAL TOOLS

INFRASTRUCTURE

HDP

Governance & Integration

Data Access

YARN

Data Management

Security

Operations

EXISTING Systems

Clickstream

Web &Social

Geolocation

Sensor & Machine

Server Logs

Unstructured

Knowledge Check

HORTONWORKS UNIVERSITY

Questions

1. Match the following components with the architectural categories of Data Management, Data Access, Governance & Integration, Security, and Operations
   - Ambari
   - HBase
   - HDFS
   - Sqoop
   - Ranger

Questions

1. Match the following components with the architectural categories of Data Management, Data Access, Governance & Integration, Security, and Operations
   - Ambari
   - HBase
   - HDFS
   - Sqoop
   - Ranger
2. True/False? The number of master nodes grows in proportion to the number of workers.

Questions

1. Match the following components with the architectural categories of Data Management, Data Access, Governance & Integration, Security, and Operations
   - Ambari
   - HBase
   - HDFS
   - Sqoop
   - Ranger
2. True/False? The number of master nodes grows in proportion to the number of workers.
3. List a few types of data sources that are new to most organizations.

Questions

1. Match the following components with the architectural categories of Data Management, Data Access, Governance & Integration, Security, and Operations
   - Ambari
   - HBase
   - HDFS
   - Sqoop
   - Ranger
2. True/False? The number of master nodes grows in proportion to the number of workers.
3. List a few types of data sources that are new to most organizations.
4. True/False? Hadoop’s goal is to displace all existing data systems.

Summary

Summary

- Hadoop ecosystem frameworks fall into the following five categories:
  - Data Management
  - Data Access
  - Governance & Integration
  - Security
  - Operations
- Primary server stereotypes are:
  - Master nodes
  - Worker nodes
- Hadoop complements existing systems and is the foundation of Connected Data Platforms

Demo: Ambari Overview
Or
Lab: Starting an HDP Cluster