Introducing Hadoop, Hortonworks, and Big Data

Lesson 2

Learning Objectives

- After you complete this lesson you should be able to:
  - Describe three categories of data
  - Describe three characteristics of Big Data
  - Describe five kinds of data commonly found in Hadoop
  - Recognize use cases for Hadoop
  - Distinguish between Hadoop and the Hortonworks Data Platform
  - Recall that Hadoop is comprised of multiple Apache projects
  - List benefits of using HDP

Consider These Scenarios

- Researchers at a hospital collect billions of data points from patient sensors. How do they analyze this data in order to improve diagnostics and patient care?
- A large online travel company collects millions of log entries. How do they store and analyze these log entries in order to improve the hotel rankings they provide to their customers?
- A nationwide trucking company collects billons of data points from sensors on their trucks. How do they analyze this data in order save fuel costs?

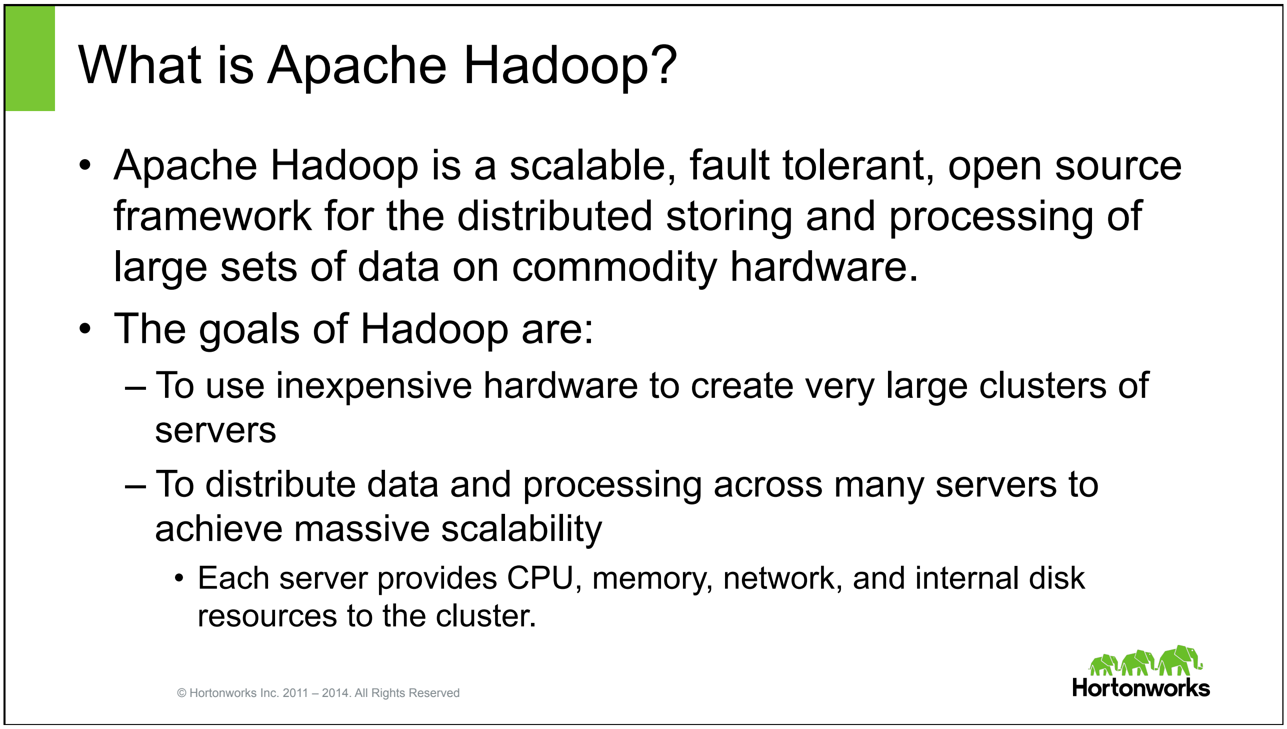
Hortonworks

The Solution is Hadoop, so YAHOO!

- Many enterprises have already discovered this solution (including Yahoo!).

UNITEDHEALTH GROUP
Georgia-Pacific
salesforce.com
at&t
FLS connect
DISCOVER
Bank of America
GAP
think FINANCE
WELLS FARGO
PROGRESSIVE
CISCO
SAMSUNG
GE Healthcare
BARNES&NOBLE BOOKSELLERS
CardinalHealth
comcast
CDW
TRUECar
UNION PACIFIC
Trustwave
OVERSEE.NET
inmar
O.C. TANNER
24/7 MEDIA
TAGGED
ENTRAVISION COMMUNICATIONS
ZirMed
move
PayPal
compete
amdocs
Spotify
Western Digital
ebay

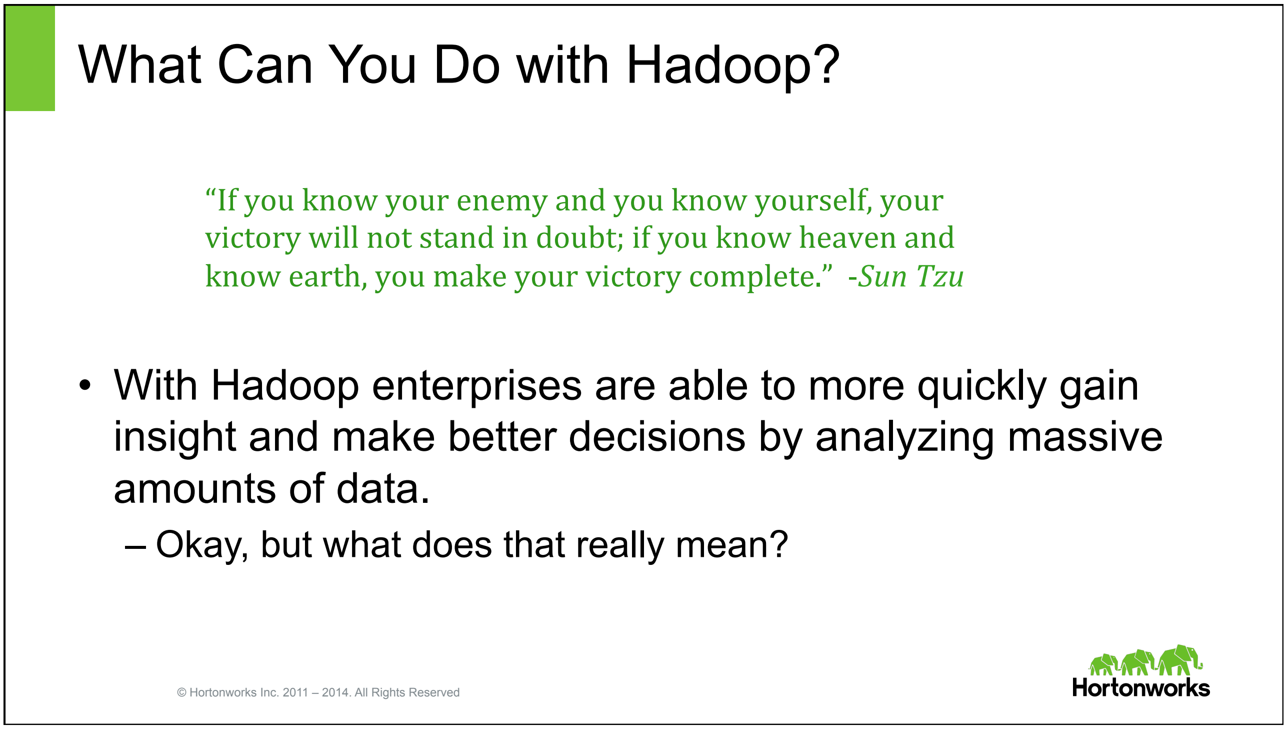
Hortonworks



What is Apache Hadoop?

- Apache Hadoop is a scalable, fault tolerant, open source framework for the distributed storing and processing of large sets of data on commodity hardware.
- The goals of Hadoop are:
  - To use inexpensive hardware to create very large clusters of servers
  - To distribute data and processing across many servers to achieve massive scalability
    - Each server provides CPU, memory, network, and internal disk resources to the cluster.

Hortonworks

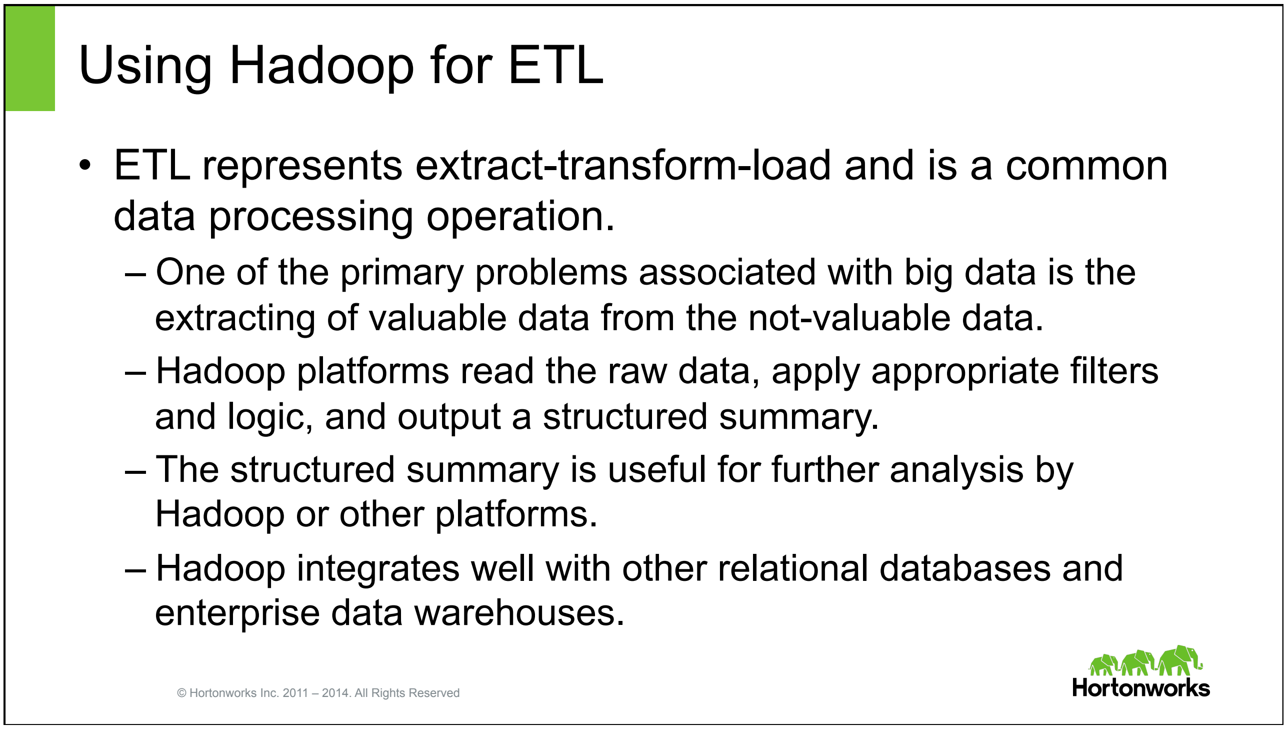


What Can You Do with Hadoop?

> "If you know your enemy and you know yourself, your victory will not stand in doubt; if you know heaven and know earth, you make your victory complete." -*Sun Tzu*

- With Hadoop enterprises are able to more quickly gain insight and make better decisions by analyzing massive amounts of data.
  - Okay, but what does that really mean?

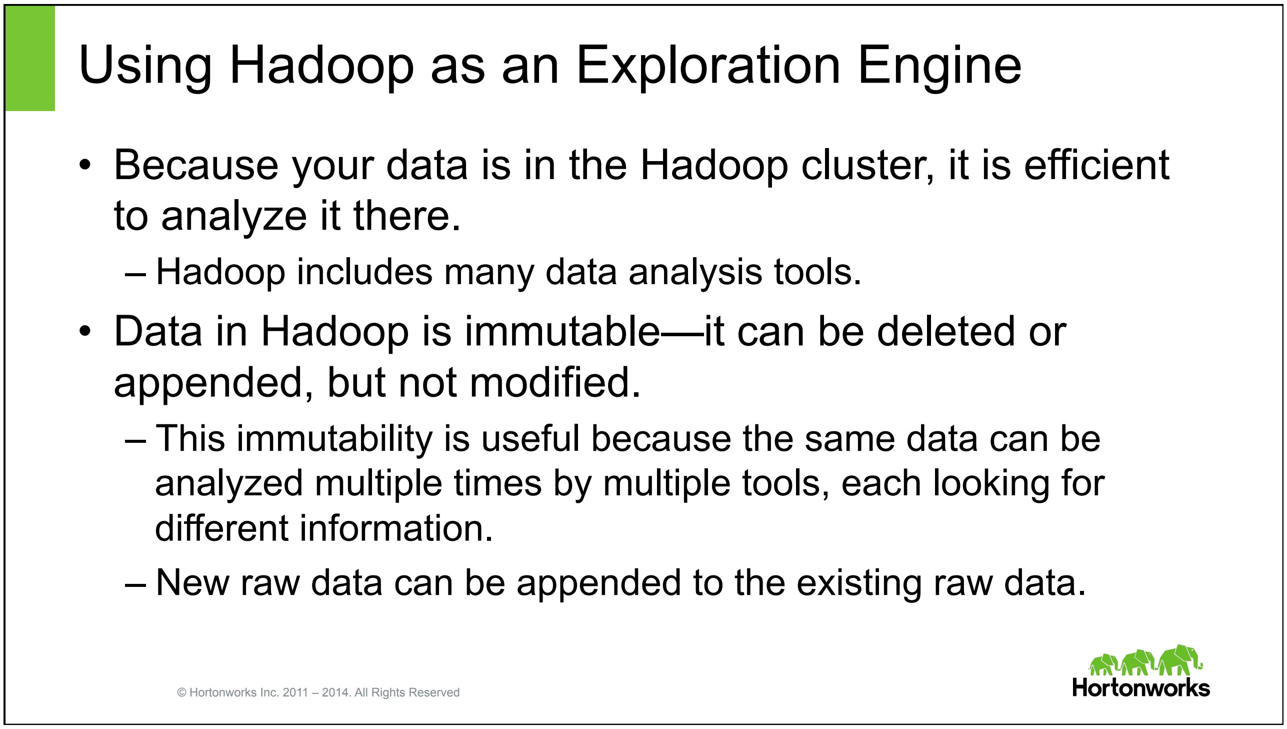
Hortonworks



Using Hadoop for ETL

- ETL represents extract-transform-load and is a common data processing operation.
  - One of the primary problems associated with big data is the extracting of valuable data from the not-valuable data.
  - Hadoop platforms read the raw data, apply appropriate filters and logic, and output a structured summary.
  - The structured summary is useful for further analysis by Hadoop or other platforms.
  - Hadoop integrates well with other relational databases and enterprise data warehouses.

Hortonworks



Using Hadoop as an Exploration Engine

- Because your data is in the Hadoop cluster, it is efficient to analyze it there.
  - Hadoop includes many data analysis tools.
- Data in Hadoop is immutable—it can be deleted or appended, but not modified.
  - This immutability is useful because the same data can be analyzed multiple times by multiple tools, each looking for different information.
  - New raw data can be appended to the existing raw data.

Hortonworks

But How Does Hadoop Really Work?

- To understand and recognize the benefits of Hadoop you must first understand data, and specifically *big* data.
- So let us talk about data.
- There are three categories of data:
  - Structured
  - Semi-structured
  - Unstructured

Hortonworks

Structured Data

- Structured data resides in a fixed field within a record or file.
  - An example would be data contained in a relational database or a spreadsheet.
- Structured data depends on creating a data model called a schema.
  - The schema defines the types of data that will be recorded and how that data will be stored, accessed, and processed.
  - This includes defining the fields of data to be stored along with the type of data in each field.
    - Types include strings, integers, floating point numbers, dates, and others.

Hortonworks

Advantages of Structured Data

- Structured data is:
  - Easily entered, stored, queried, and analyzed
  - Often managed using a structured query language (SQL)

Hortonworks

Semi-Structured Data

- Data that does not conform to a formal data model defined in a schema
- Data that does contain tags or other markers
  - These separate semantic elements and enforce hierarchies of records and fields within the data.
- Examples include XML, HTML, and Java Script Object Notation (JSON) files.
- It is sometimes called self-describing data.

Hortonworks

Unstructured Data

- Data that does not reside in fixed fields or records
- There is no data model, contained in a schema or tags, that defines how to store, access, or process the data.
- Examples include word processing documents, videos, photos, audio files, presentations, Web pages, and many other kinds of business documents.
  - These examples may have an internal structure, but they do not have schema, or internal tags or metadata, describing their fields of data.

Hortonworks

Disadvantages of Unstructured Data

- Unstructured data has irregularities and ambiguities that make it difficult to understand and analyze using traditional computer programs.
  - This data is often ignored or deleted which limits its business value.
  - Experts estimate that 80-90 percent of the data in any enterprise is unstructured.
    - Hadoop’s primary contribution has been the capability to extract value from this unstructured data.

Hortonworks

Unstructured Data and Hadoop

- Hadoop excels at analyzing unstructured data.
  - In some cases Hadoop tools can impose a structure on unstructured data so that it can be analyzed.
    - The structure is layered over the raw data and does not change the raw data.
    - Remember, data in Hadoop is immutable.
  - In other cases Hadoop tools transform the data to create structure.
    - Transformed data is saved as new data.
    - The transformed data is analyzed by Hadoop or other platforms.

Common Types of Data in Hadoop

- There are five types of data commonly found in Hadoop.
  - Sentiment data
  - Clickstream data
  - Sensor or machine data
  - Geolocation data
  - Server log data

What Makes Data BIG DATA?

- The term Big Data comes from the computational sciences.
- It is used to describe scenarios where the volume and types of data overwhelm the tools to store and process it.
- In 2001, industry analyst Doug Laney described Big Data using the three V's:

VOLUME *Velocity* Variety

Hortonworks

Volume

- Volume refers to the amount of data being generated.
  - Gigabytes, terabytes, petabytes…
  - Many factors contribute to the increase in data volume, including:
    - Transaction-based data stored through the years
    - Unstructured data streaming in from social media
    - Increasing amounts of sensor and machine-to-machine data being collected
  - Problems related to volume include:
    - Storage costs
    - Determining relevance within large data volumes
    - How to analyze data quickly to maximize business value

Hortonworks

Velocity

- Velocity refers the rate at which new data is generated.
  - Megabytes per second, gigabytes per second…
  - Data is streaming in at unprecedented speed and must be dealt with in a timely manner in order to extract the maximum value.
    - Sources include logs, social media, RFID tags, sensors, and smart metering
  - Problems related to velocity include:
    - Reacting quickly enough to benefit from the data
    - Inconsistent data flows with periodic peaks
      - Daily, seasonal, and event-triggered peak data loads
        - » A new trend in social media.

Variety

- Variety refers to the number of types of data being generated.
  - Varieties of data include:
    - Structured data in traditional databases
    - Semi-structured data like XML or JSON files
    - Unstructured text documents, email, video, audio, stock ticker data, and financial transactions
  - Problems related to variety include:
    - How to gather, link, match, cleanse, and transform data across systems
    - How to connect and correlate data relationships and hierarchies to extract business value

Hadoop Was Designed for Big Data

> "Big Data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." – *Gartner*

- Hadoop is the framework that provides a cost-effective, innovative form of information processing used to enhance business insight and decision making.

Hortonworks

So How Does Hortonworks Fit Into All of This?

- Hortonworks does Hadoop.
  - In fact, it does it very well.
- Hortonworks:
  - Was founded in 2011 by 24 engineers from the original Yahoo! Hadoop development and operations team
  - Has amassed more Hadoop experience under one organization than anyone else
  - Has team members who are active participants and leaders in Hadoop development
  - Has years of experience in Hadoop operations

Hortonworks

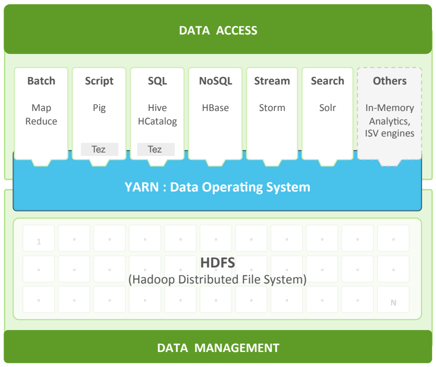
But Wait, There is More…

- Hortonworks is also:
  - Responsible for approximately 50 percent of core code base advances used to deliver Apache Hadoop as an enterprise data platform
  - Partnered with trusted data center companies like Microsoft, Red Hat, SAP, Teradata, Rackspace, and many more
  - Building Hadoop with the enterprise in mind, all tested and certified with real-world vigor in one of the world's largest Hadoop clusters
  - A source of world-class enterprise support, consulting, and training

Hortonworks

Hadoop and the Hortonworks Data Platform (HDP)

- What is the difference between Hadoop and HDP?
- To answer, we need more information about Hadoop.
- Hadoop is actually a framework that includes many components.
  - The components are a collection of other frameworks and tools that are managed as *projects* by the Apache Software Foundation.
  - A few projects are illustrated here.



Hortonworks

Why So Many Frameworks and Tools?

- There are so many tools because each tool was created to provide a specific solution.
- Some tools are easier to use in specific situations even though some tools have functional overlap.
  - As an analogy, a screwdriver can be used as a chisel but using a chisel is easier.
  - Likewise, Apache Storm and Apache Flume can be used to ingest data and perform real-time analysis, but real-time analysis is easier with Storm.

Installing and Using Hadoop Projects

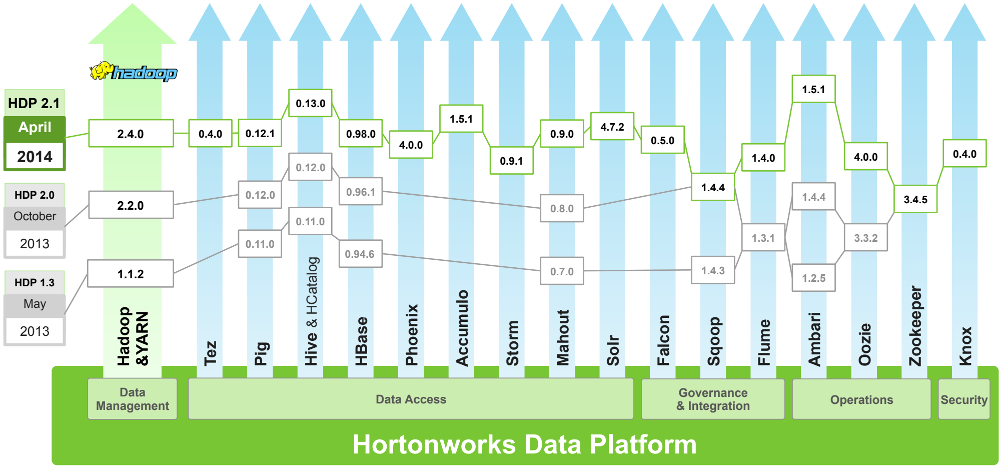
- It is important to understand that each project is developed independently and has its own release schedule.
- Any organization is free to download, install, and test any version of a project from the Apache Software Foundation.
  - However, it takes a tremendous amount of skill and testing to find the right combination of project versions.
  - HDP exists to make it possible to avoid this scenario.

Getting Back to HDP – What is it?

- HDP is an enterprise version of Hadoop distributed by Hortonworks.
  - It includes a single installation utility that installs most of the Apache Hadoop software.
  - The benefit is that it has gone through rigorous system, function, and regression testing to ensure that versions of any projects included in the distribution work seamlessly together in a secure and reliable manner.
    - It is important for an enterprise version of Hadoop to use the best combination of the most stable, reliable, secure, and current projects.

Hortonworks

HDP and Apache Hadoop Project Versions



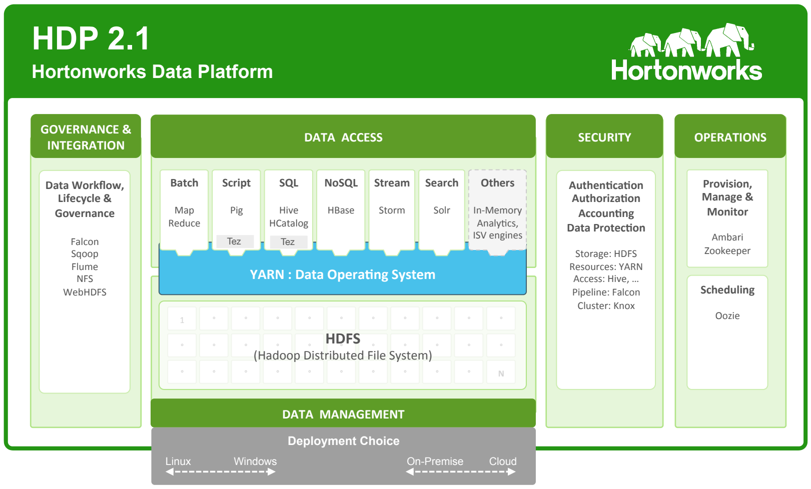
Hortonworks

HDP is 100 Percent Pure Hadoop

- HDP contains no proprietary software add-ons or extensions.
  - This ensures that you receive 100 percent Apache Hadoop software with no possibility of vendor lock-in.
  - You benefit from the latest enhancements and fixes developed by a worldwide set of developers.
    - You do not have to wait for a single company to add the enhancements and fixes to their proprietary software.
  - Other software often runs along side of Hadoop and might not work properly with proprietary extensions and features.

Hortonworks

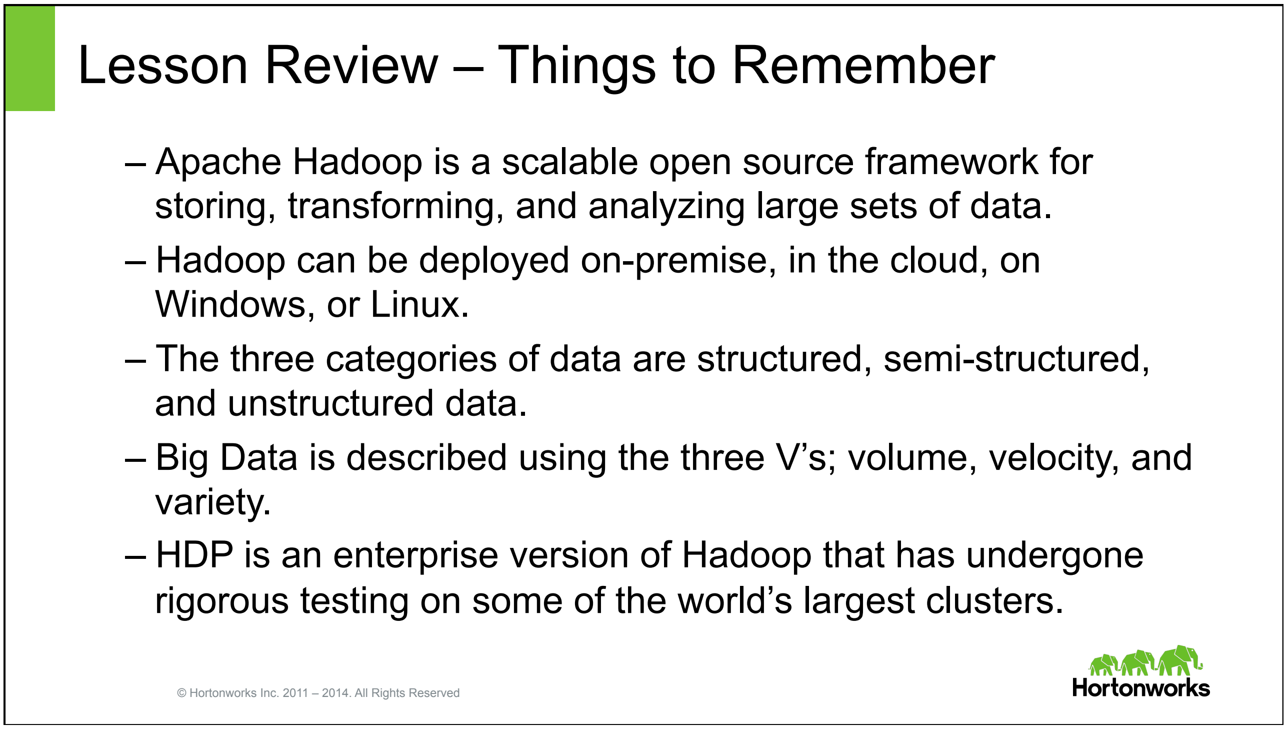
HDP Delivers a Comprehensive Data Management Platform



HDP includes projects for:

- Core Hadoop (HDFS and YARN)
- Data access and processing
- Data governance and integration
- Security
- Managing Hadoop operations

Hortonworks



Lesson Review – Things to Remember

- Apache Hadoop is a scalable open source framework for storing, transforming, and analyzing large sets of data.
- Hadoop can be deployed on-premise, in the cloud, on Windows, or Linux.
- The three categories of data are structured, semi-structured, and unstructured data.
- Big Data is described using the three V's; volume, velocity, and variety.
- HDP is an enterprise version of Hadoop that has undergone rigorous testing on some of the world's largest clusters.

Hortonworks