Creating Your First Big Data Hadoop Cluster Using Cloudera CDH

THE CASE FOR BIG DATA, HADOOP, & CLOUDERA

Xavier Morera

PASSIONATE ABOUT TEACHING

@xmorera www.xaviermorera.com

Objectives of This Training

Big Data, Hadoop & Cloudera

Fast Track to Big Data with QuickStart VM

Play with Big Data "Ask Bigger Questions"

Set up a Virtualization Environment

Build Your First Cluster using Cloudera CDH

Take Your First Steps Working with a Cluster

And This Is My Promise to You

The Case for Big Data

Information explosion

- Generating data faster than ever
- Human created data has grown
- But machine generated data exploded

Data has valuable applications

But we need to store, process and analyze

Limitations with existing technologies

Something else, totally different was needed

To process and make sense of all this data

What’s the Definition of Big Data?

Defining Properties of Big Data

Volume

Velocity

Variety

Value!

Big Data for the Relational Mind

Relational Databases	NoSQL & Big Data
Structured data	Unstructured data
Relational, schema oriented	Heterogeneous sources
Normalized	Stored as is, schema less
Scaled vertically (bigger boxes)	Scale horizontally (many boxes)
Data brought to computation	Computation brought to data
Data archived or “thrown away”	Larger data sets processed for longer
Limitations	Scale

How Did It Work Before?

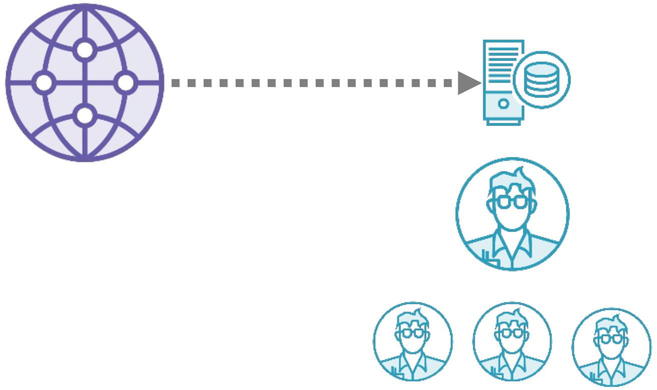
You Had a Bunch of Data

Buy Biggest Box You Can Afford

And Most Powerful Database

Get Answers to Your Questions

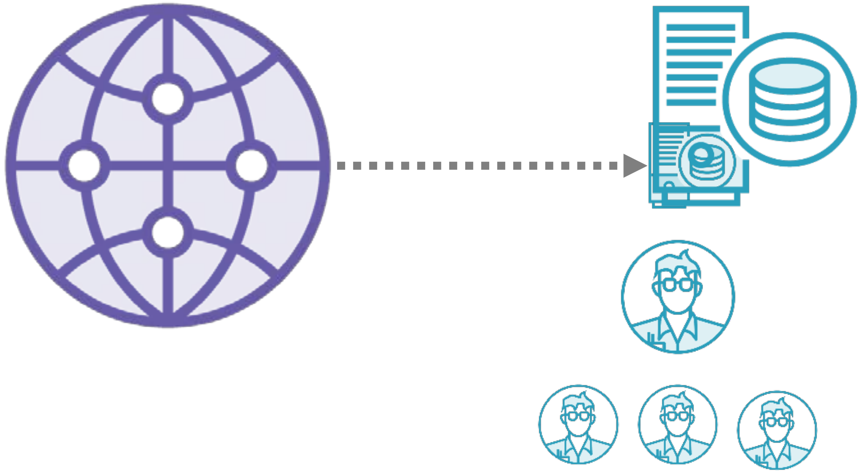
Let’s Talk About Scale



Let’s solve a problem we face every day

Search the internet!

Let’s Talk About Scale



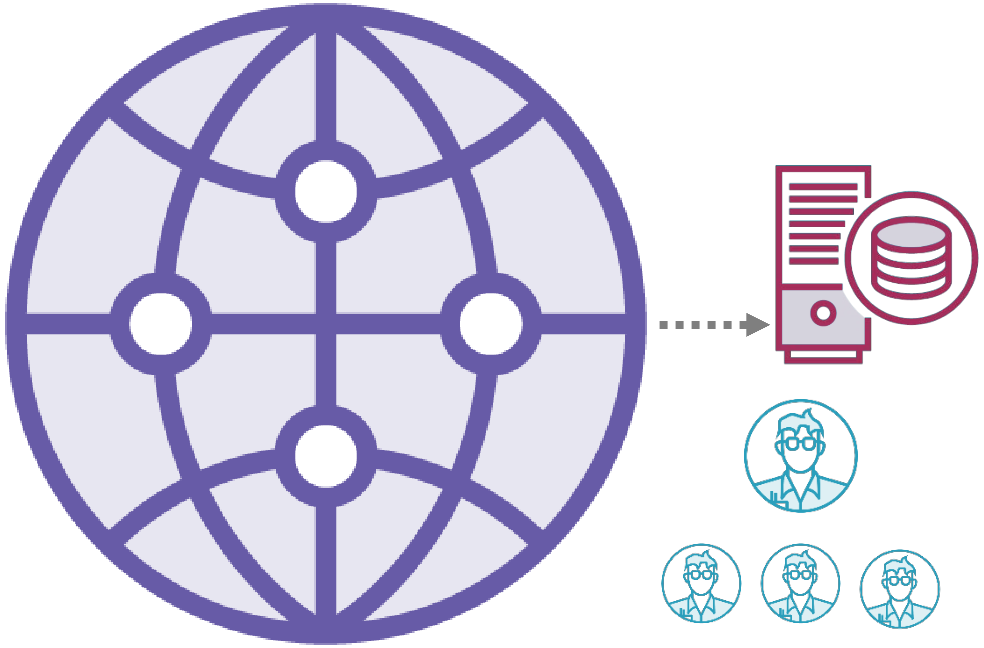
And as the amount of data grows

You simply buy a bigger machine

And you have your “Data Temple”

$$$

Let’s Talk About Scale



But eventually it is too much data

Scaling up is not an option

How Do You Solve a Big Data Problem?

Distributed Computing!

Welcome to a Big Data World

Perfect Timing

New Computing Platform Invented by Google

Commodity Hardware Available Cheaply

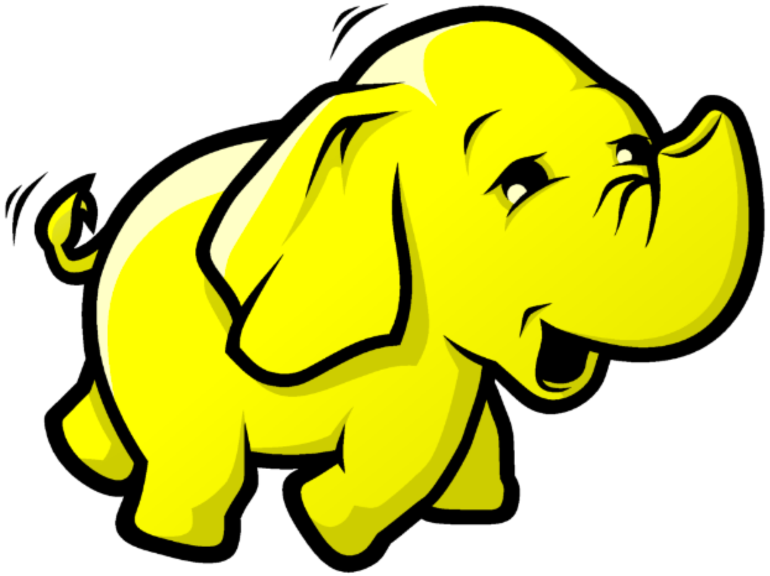
Data Explosion

Big Data

Big Data Is the New Competitive Advantage

Big Data Is Hadoop

The Case for Hadoop



Part of Apache Software Foundation

New Type of Data Platform

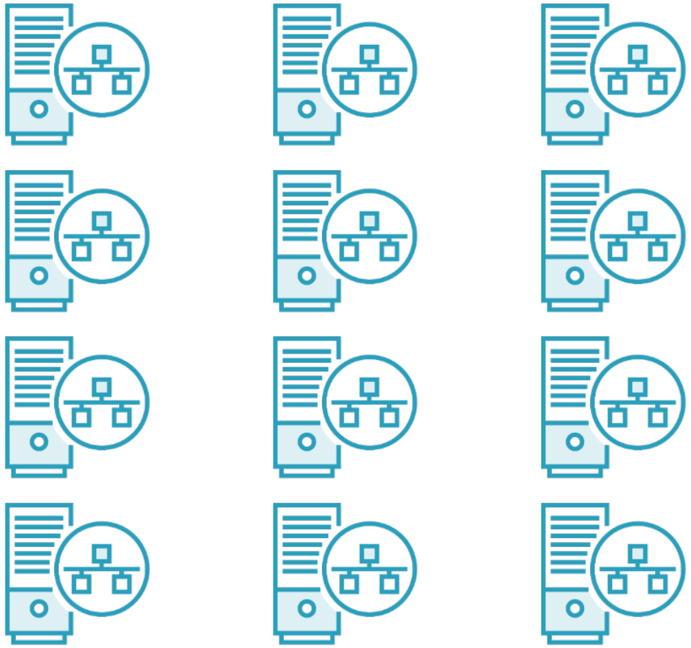
- Store virtually unlimited data
- Process with different frameworks

Tested in production

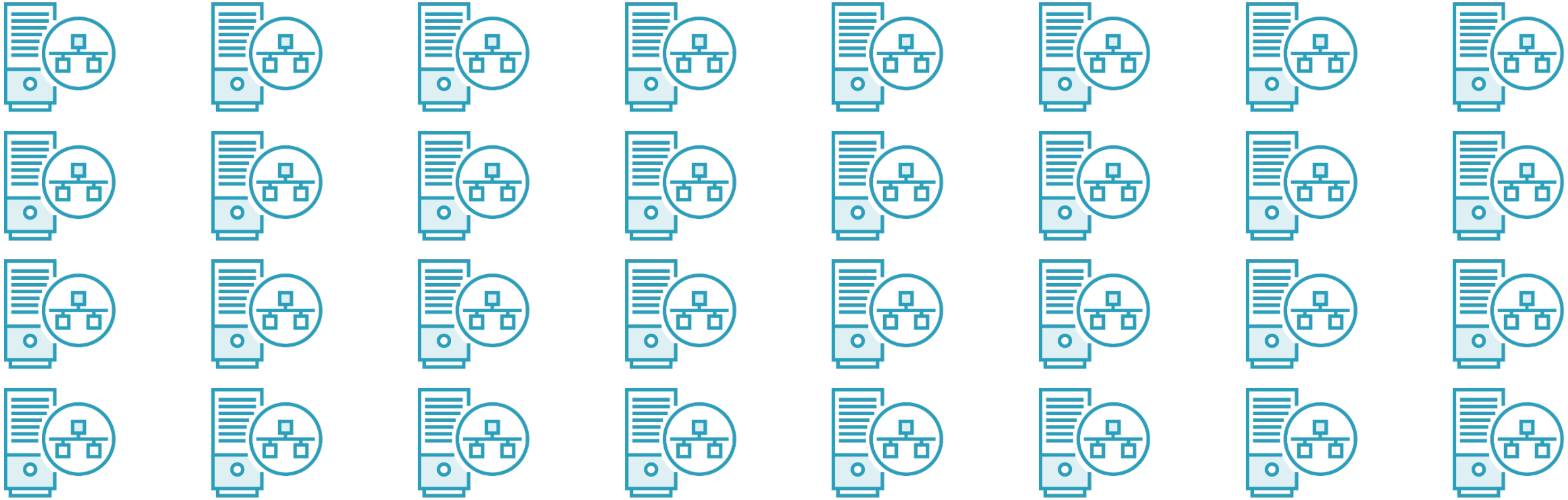
- Sensitive & critical data
- Huge amounts of data
- Used by some of largest companies

Cost Effective

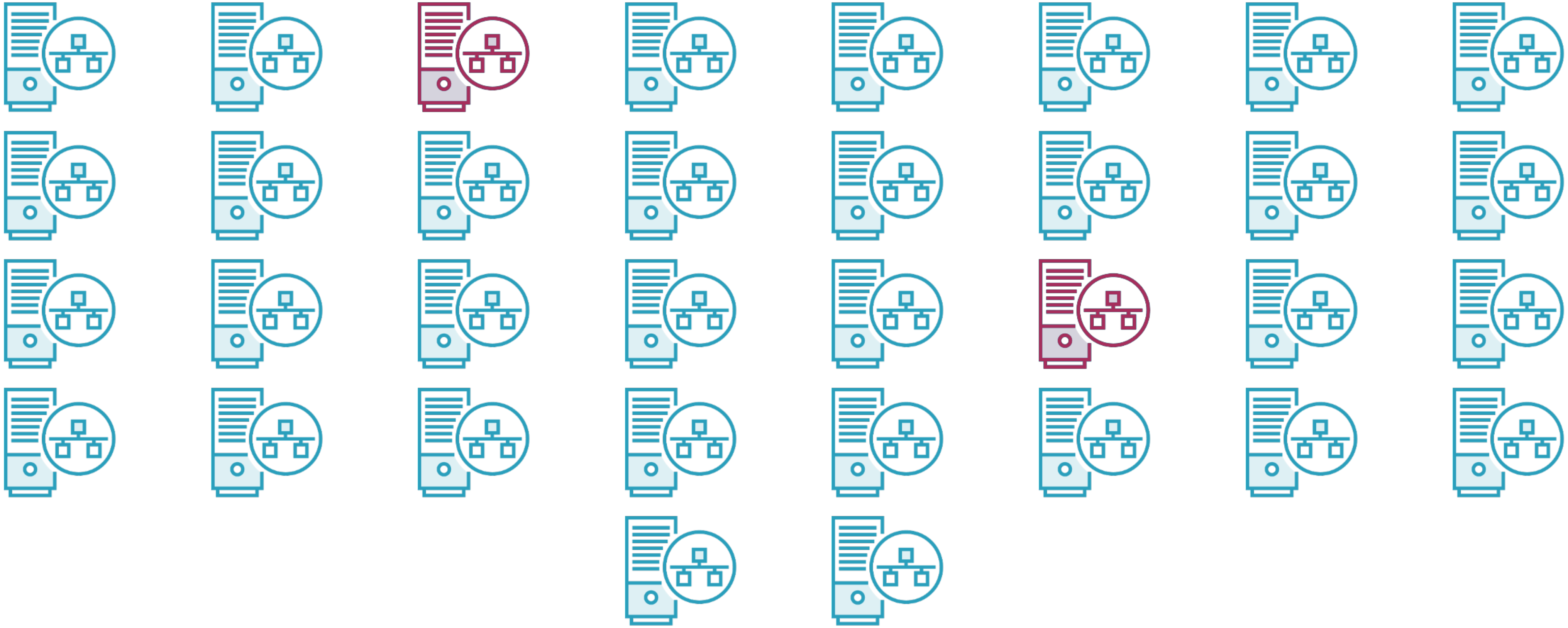
vs.



Scalability & Speed: Parallelism at Its Best



Resiliency for the Win



Open Source

$0.00

Cost of Getting Software

Something Missing?

Contribute!

Vendor Lock-in

Best of All

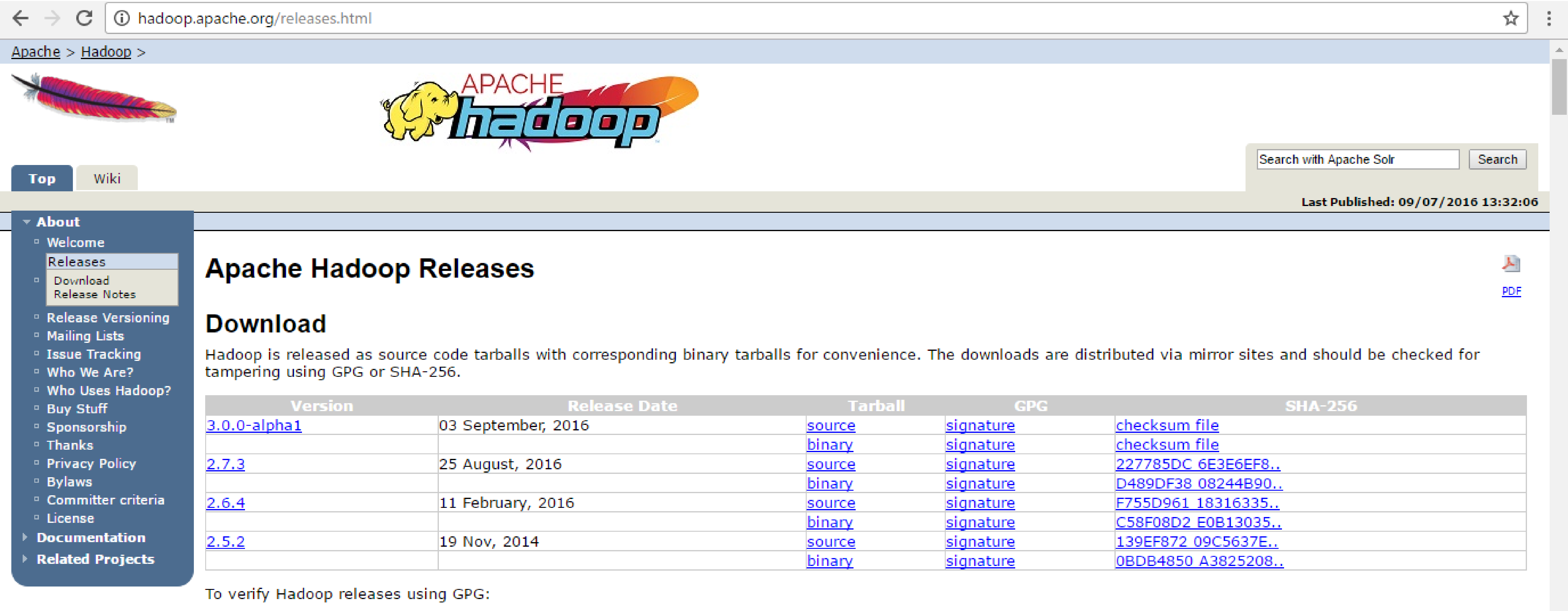
No Vendor Lock-in

“Hadoop Is the Most Disruptive Technology in Data Management in Our Lifetime”

A Very Smart Engineer Turned Cloudera Product Owner Told Me

How Do I Get Hadoop?

Or as I call it: “Hadoop the Hard Way”



Last Published: 09/07/2016 13:32:06

Apache Hadoop Releases

Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

Version	Release Date	Tarball	GPG	SHA-256
3.0.0-alpha1	03 September, 2016	source	signature	checksum file
		binary	signature	checksum file
2.7.3	25 August, 2016	source	signature	227785DC 6E3E6EF8..
		binary	signature	D489DF38 08244B90..
2.6.4	11 February, 2016	source	signature	F755D961 18316335..
		binary	signature	C58F08D2 E0B13035..
2.5.2	19 Nov, 2014	source	signature	139EF872 09C5637E..
		binary	signature	0BDB4850 A3825208..

To verify Hadoop releases using GPG:

Some “Assembly” Required

But there are a few gotcha's

Multiple versions

- Easy to make mistakes

Too many projects working together

- Where do I start?

Too many manual steps

- Thus error prone

Many potential hair pulling scenarios

Easy to Feel Overwhelmed When Starting

And This Is Where Cloudera Comes In!

The Case for Cloudera

Vision + Knowledge + Perfect timing

Group of smart people had a vision

- Saw an unexploited opportunity

Big Data for the masses

Embrace Open Source and contribute

Very successful advancing Big Data

Cloudera’s Added Value

Core is Open Source

Production Tools
(Proprietary)

Services & Support

A Little Bit About cloudera

2008

Founded

Mike Olson
Christophe Bisciglia
Amr Awadallah
Jeff Hammerbacher

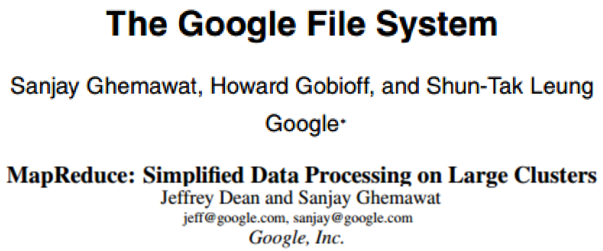
Founders

SLEEPYCAT SOFTWARE
Makers of Berkeley DB
ORACLE
facebook
YAHOO!
Google

Who Came From

hadoop

Founded a Company Around



The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*

MapReduce: Simplified Data Processing on Large Clusters
Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.

GFS + MapReduce Papers by Google

~$0

Market Value at the Time?

More About cloudera

Multi Billion Dollar Company

And They Became

Doug Cutting
Tom White
Yonik Seely
And Many More

Some of the Best

50%+ of Their Engineering Donated to Projects

Major Contributor to Open Source

Adopted by
Fortune 100
+
2,500+ Partners

Ecosystem

Software, Services & Support

They Offer

CDH

They Created

Cloudera's Distribution, Including Hadoop

Known as CDH

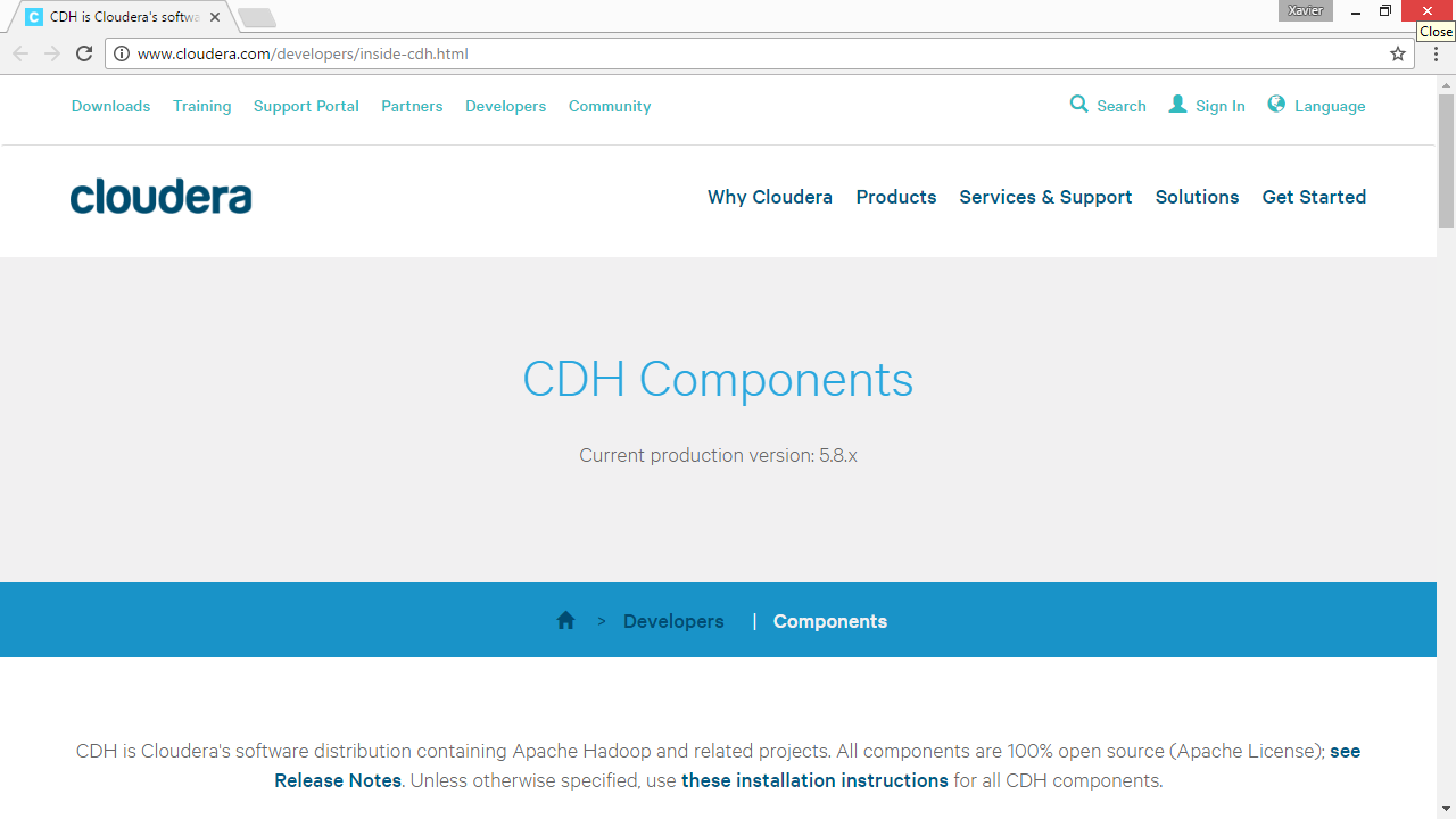
100% Open Source

Distribution of Hadoop

Enterprise ready

- Widely deployed and tested

Integrates key Hadoop projects



CDH Components

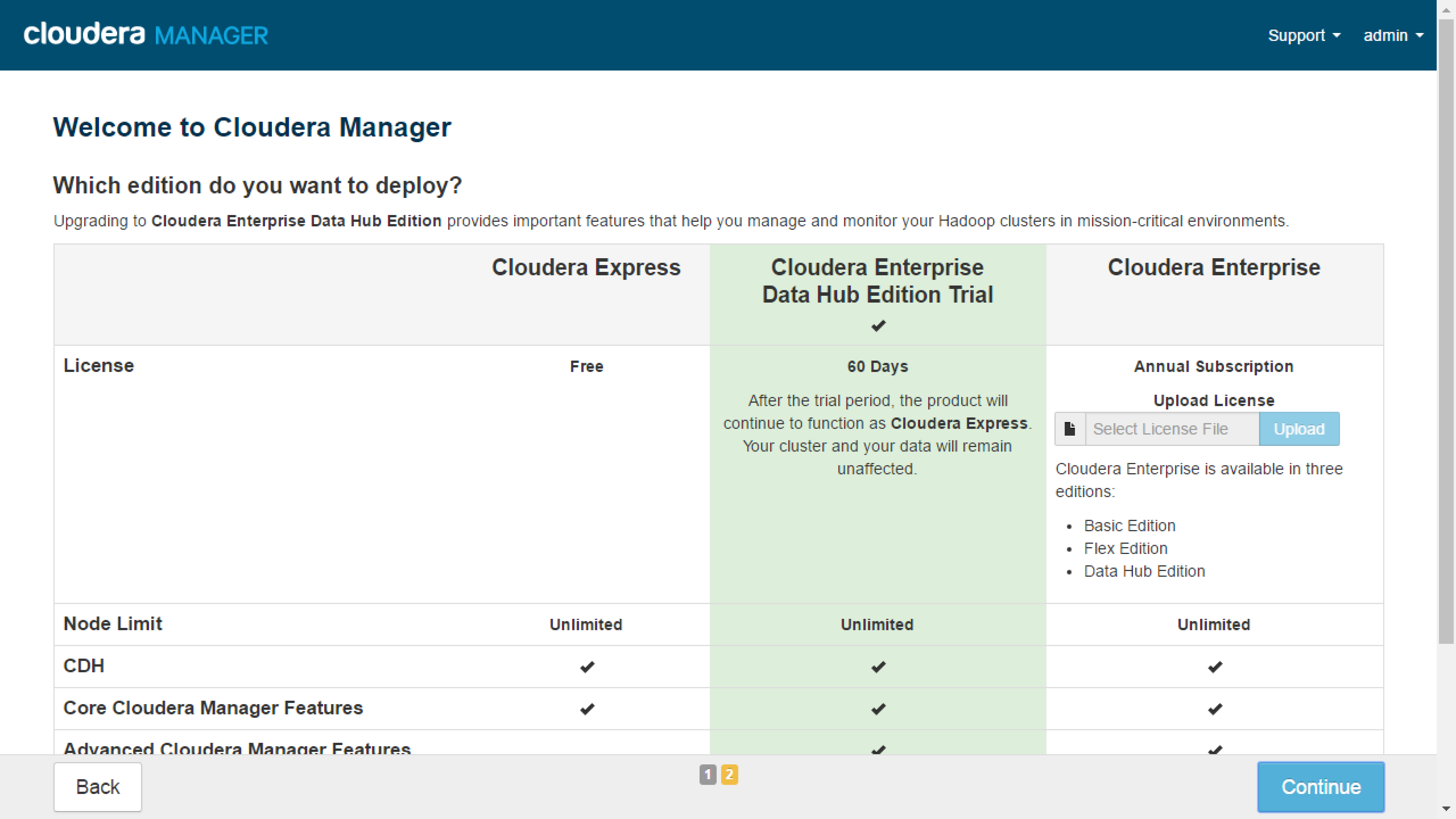
Current production version: 5.8.x

CDH is Cloudera's software distribution containing Apache Hadoop and related projects. All components are 100% open source (Apache License); **see Release Notes**. Unless otherwise specified, use **these installation instructions** for all CDH components.

Which Cloudera Version for Me?

Several different versions

You need to decide which one you will use



Welcome to Cloudera Manager

Which edition do you want to deploy?

Upgrading to **Cloudera Enterprise Data Hub Edition** provides important features that help you manage and monitor your Hadoop clusters in mission-critical environments.

	Cloudera Express	Cloudera Enterprise Data Hub Edition Trial ✓	Cloudera Enterprise
License	Free	60 Days After the trial period, the product will continue to function as **Cloudera Express**. Your cluster and your data will remain unaffected.	Annual Subscription Upload License Select License File Upload Cloudera Enterprise is available in three editions: • Basic Edition • Flex Edition • Data Hub Edition
Node Limit	Unlimited	Unlimited	Unlimited
CDH	✓	✓	✓
Core Cloudera Manager Features	✓	✓	✓
Advanced Cloudera Manager Features		✓	✓

Which Cloudera Version for Me?

Several different versions

You need to decide which one you will use

All versions include unlimited hosts

Cloudera Express

- Includes Cloudera Manager

Cloudera Enterprise

- Paid
- Support
- Additional features & tools

Cloudera Express

All key Apache Hadoop ecosystem projects

Cloudera Manager

- No rolling upgrades
- No backup/disaster recovery
- No LDAP/SNMP integration

No technical support from Cloudera

Everything You Need to Get Started with Hadoop (and Production)

Cloudera Enterprise

Where Cloudera helps you go to Production

Provide the support you need

Additional features & tools

Licensing

- Annual subscription per node
- Elastic pricing available

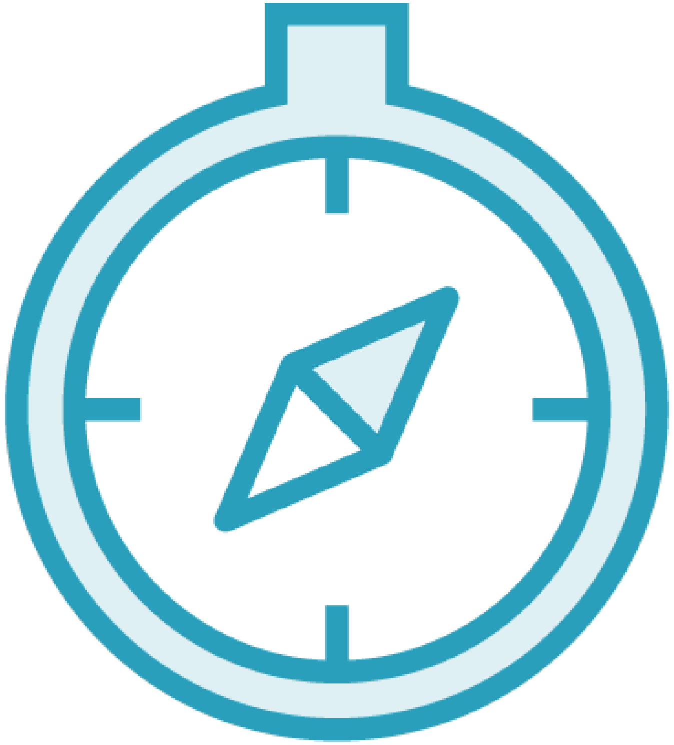
Multiple editions

- Based on products you want support for

Cloudera's got your back!

Making Hadoop
Fast, Easy, and Secure

Cloudera Navigator



One of the Enterprise tools from Cloudera

Data Governance solution for Hadoop

Key Capabilities

- Data discovery
- Continuous optimization
- Audit
- Lineage
- Metadata management
- Policy enforcement

Critical part of Cloudera Enterprise

Big Data Meets Data Governance

Cloudera Is Your Top Choice for Getting into Big Data

Takeaway

Digital revolution

Volume, Velocity, Variety & Value

Scale, resiliency, performance

Big Data, Hadoop & Cloudera

CDH

Cloudera Express or Enterprise