The Case for Hadoop

Lesson Objectives

After completing this lesson, students should be able to:

- Describe data trends of volume, velocity and variety
  - Technology threats and opportunities
- List popular use cases for Hadoop
- Discuss the importance of Open Enterprise Hadoop
  - Open
  - Central
  - Interoperable
  - Ready
- Give an overview of Connected Data Platforms powered by Hadoop

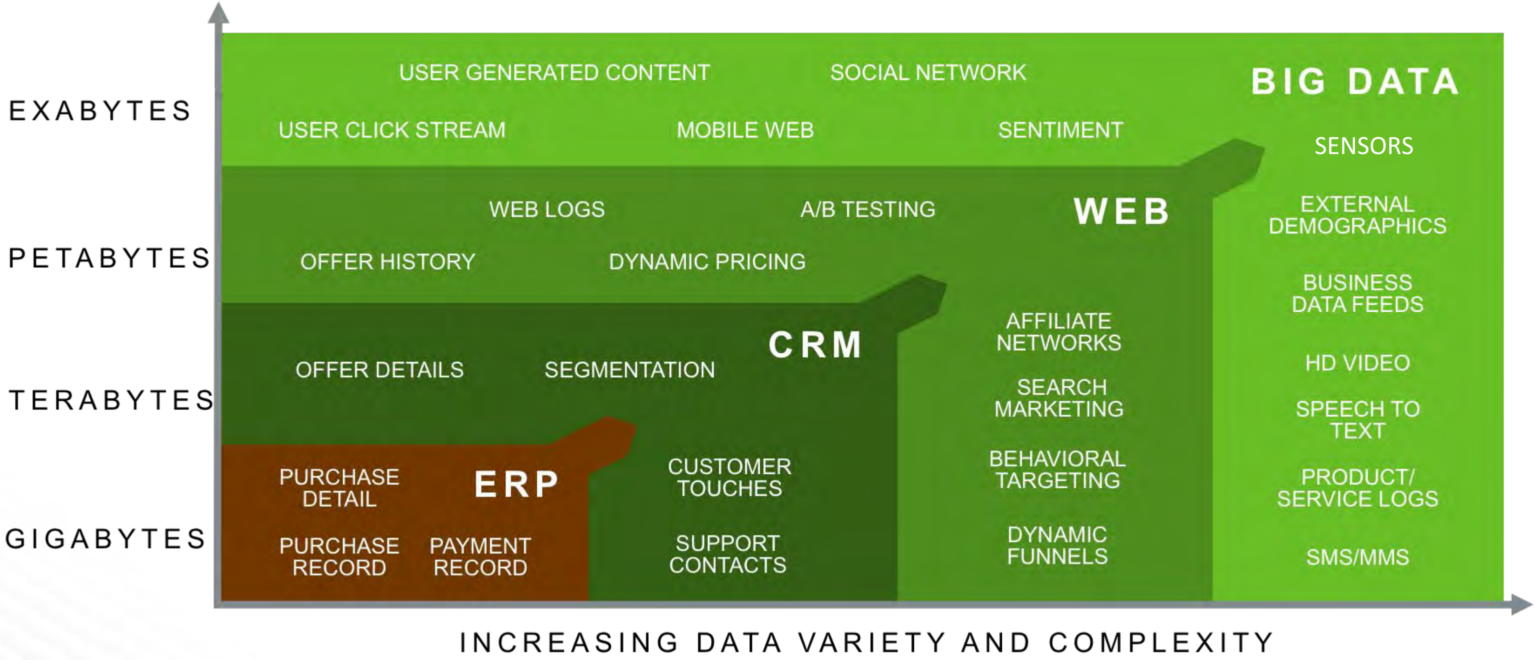
Data Trends

Popular Use Cases for Hadoop

Open Enterprise Hadoop

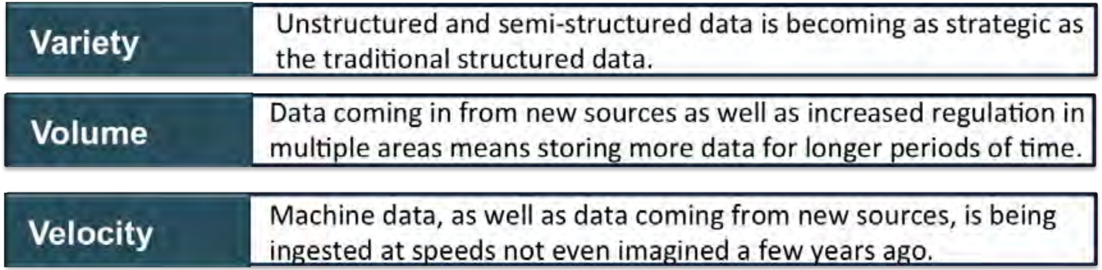
Why Hortonworks?

The 3 V's of DATA are Driving Apache Hadoop



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



Variety	Unstructured and semi-structured data is becoming as strategic as the traditional structured data.
Volume	Data coming in from new sources as well as increased regulation in multiple areas means storing more data for longer periods of time.
Velocity	Machine data, as well as data coming from new sources, is being ingested at speeds not even imagined a few years ago.

VOLUME *Velocity* Variety

Volume

Volume refers to the amount of data being generated.

- Gigabytes, terabytes, petabytes, exabytes, zettabytes …
- 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

Velocity

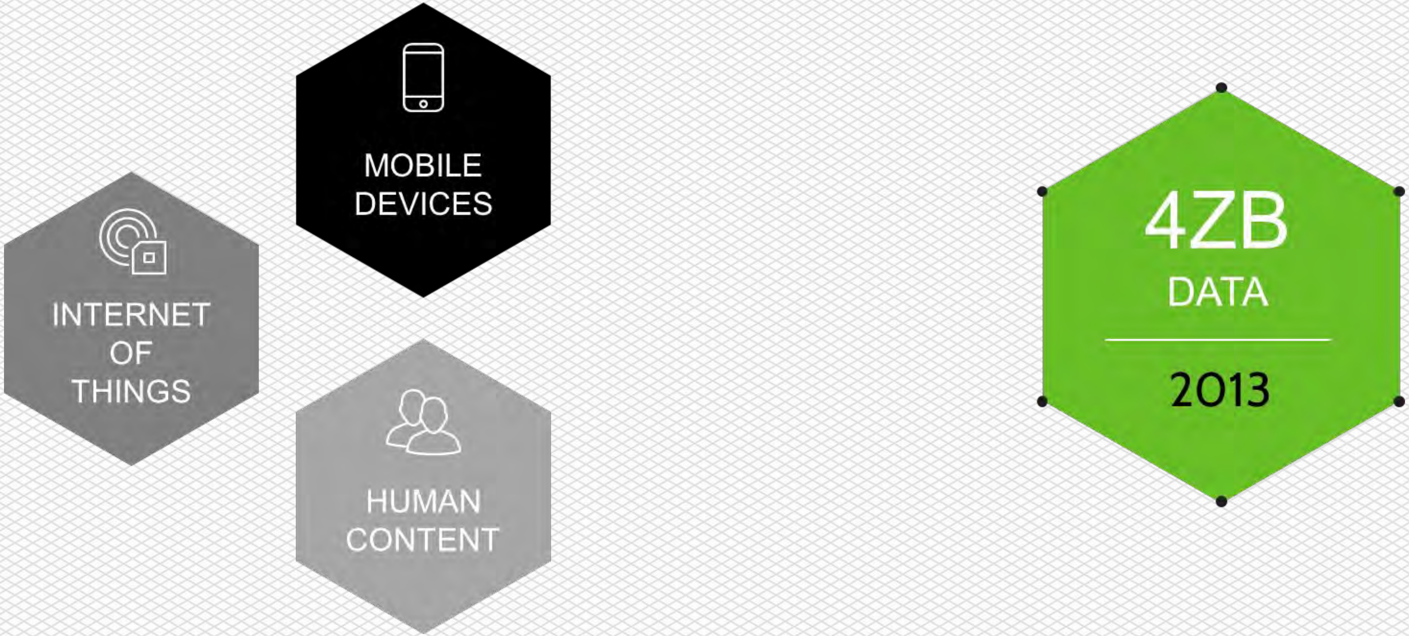
Velocity refers the rate at which new data is generated.

- 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
    - Event-triggered

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



MOBILE DEVICES

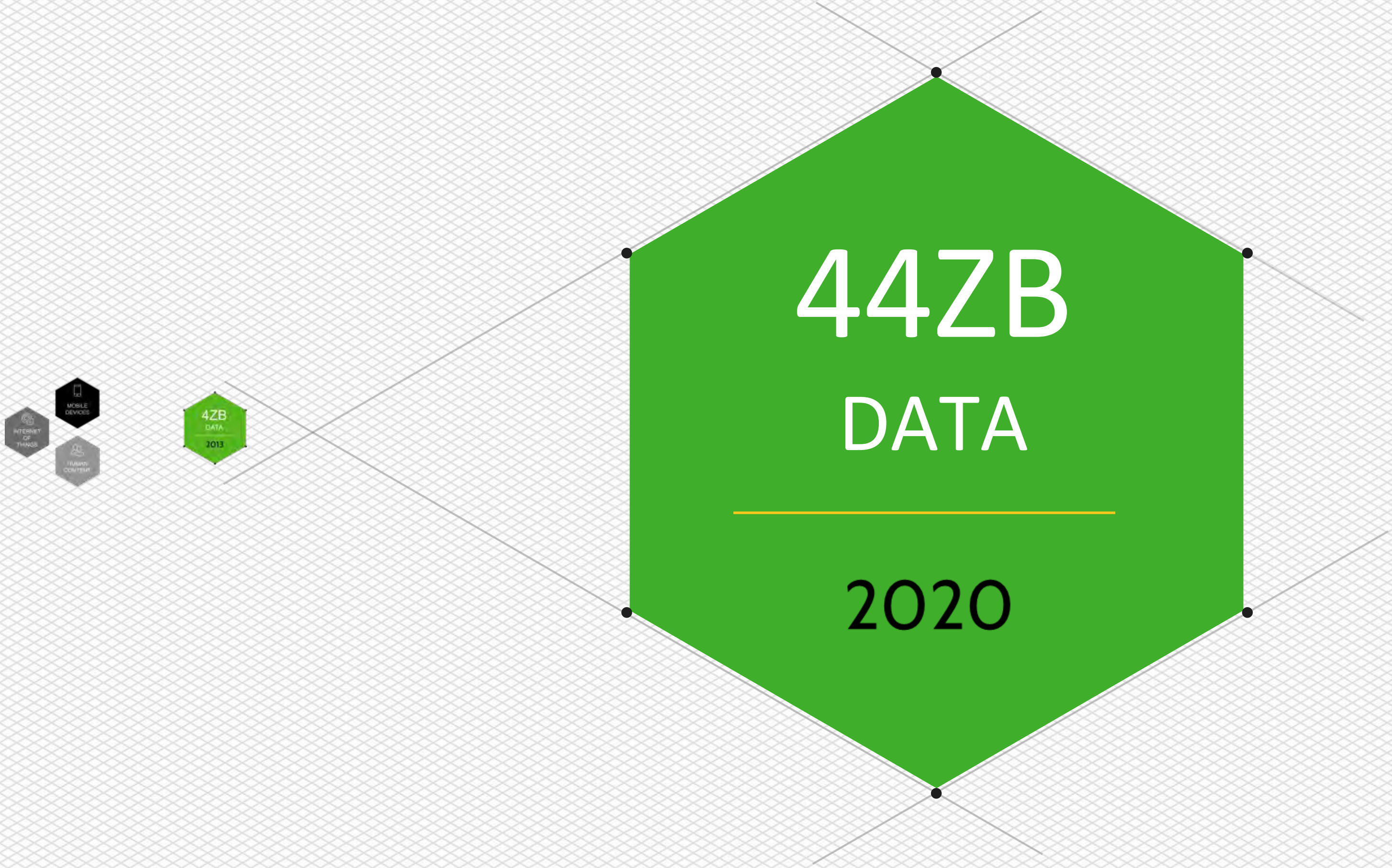
INTERNET OF THINGS

HUMAN CONTENT

4ZB DATA

2013

Source: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm



44ZB

DATA

2020

Source: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Threats

Existing data architectures make data inaccessible, incomplete, irrelevant, and expensive.

Opportunities

Apache™ Hadoop® transforms your business, making Big Data easily accessible for advanced analytic applications.

What is Apache Hadoop?

The Apache Hadoop project describes the technology as a software framework that:

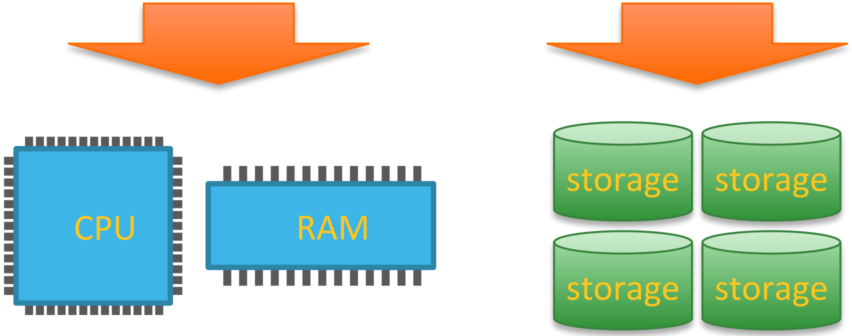
- Allows for the distributed processing of large data sets across clusters of computers using simple programming models
- Is designed to scale up from single servers to thousands of machines, each offering local computation and storage
- Does not rely on hardware to deliver high-availability, but rather the library itself is designed to detect and handle failures at the application layer
- Delivers a highly-available service on top of a cluster of computers, each of which may be prone to failures

Source: http://hadoop.apache.org

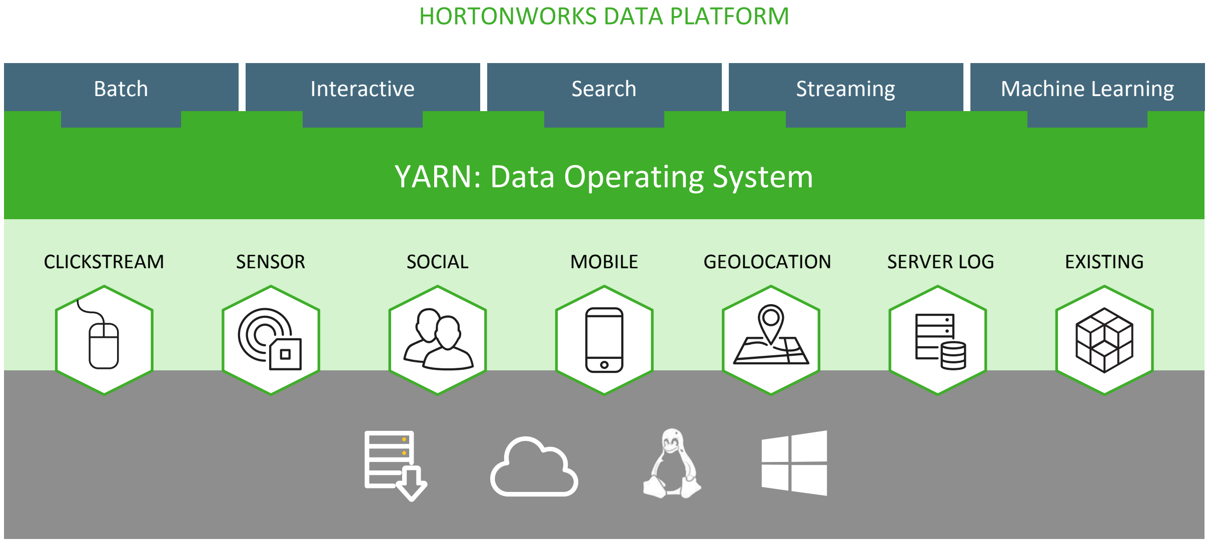
Hadoop Core = Storage + Compute

Yet Another Resource Negotiator (YARN)

Hadoop Distributed File System (HDFS)



Hortonworks Delivers Open Enterprise Hadoop

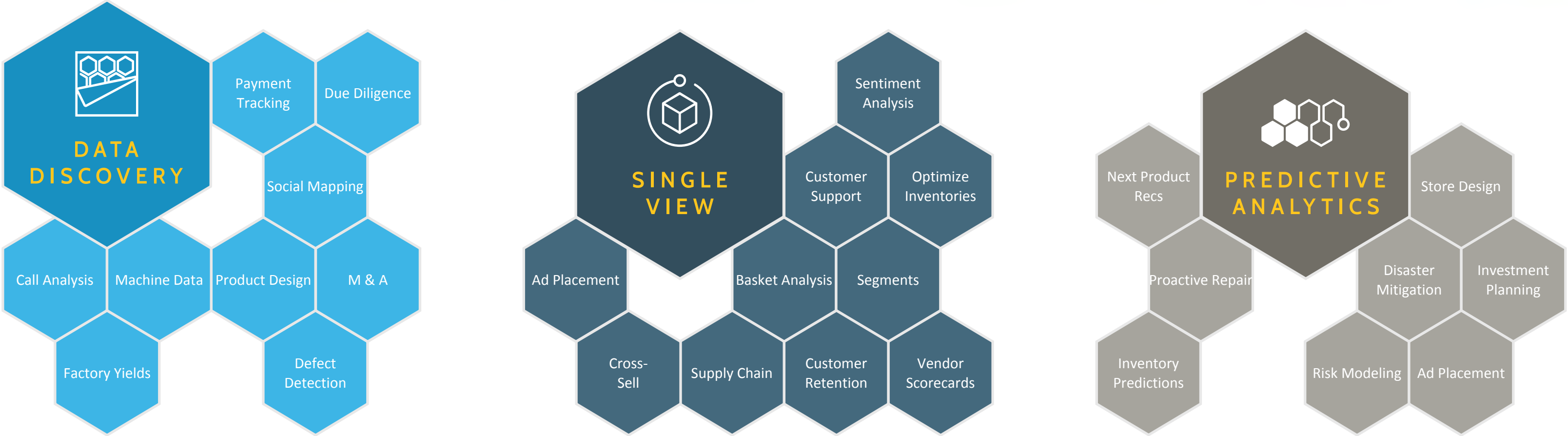


Data Trends

→ **Popular Use Cases for Hadoop**

Open Enterprise Hadoop

Why Hortonworks?

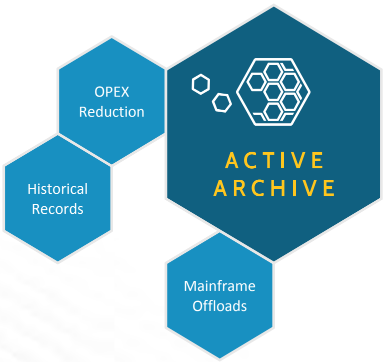


BUSINESS OUTCOMES

Business executives are driving transformational outcomes with next-generation applications that empower new uses of Big Data including: data discovery,
a single view of the customer and predictive analytics.

COST SAVINGS

IT executives are delivering substantial reductions in operating costs by modernizing their data architectures with Open Enterprise Hadoop. These cost saving innovations include active archive of cold data, offloading ETL processes and enriching existing data.

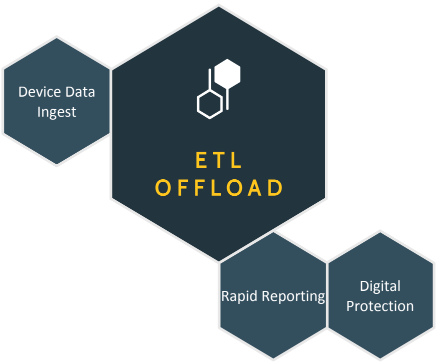


ACTIVE ARCHIVE

OPEX Reduction

Historical Records

Mainframe Offloads

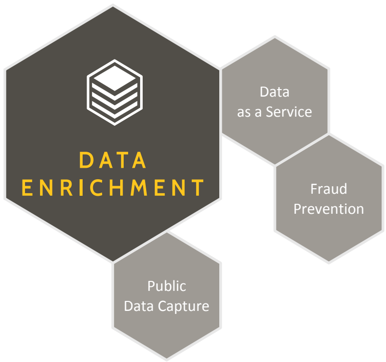


ETL OFFLOAD

Device Data Ingest

Rapid Reporting

Digital Protection

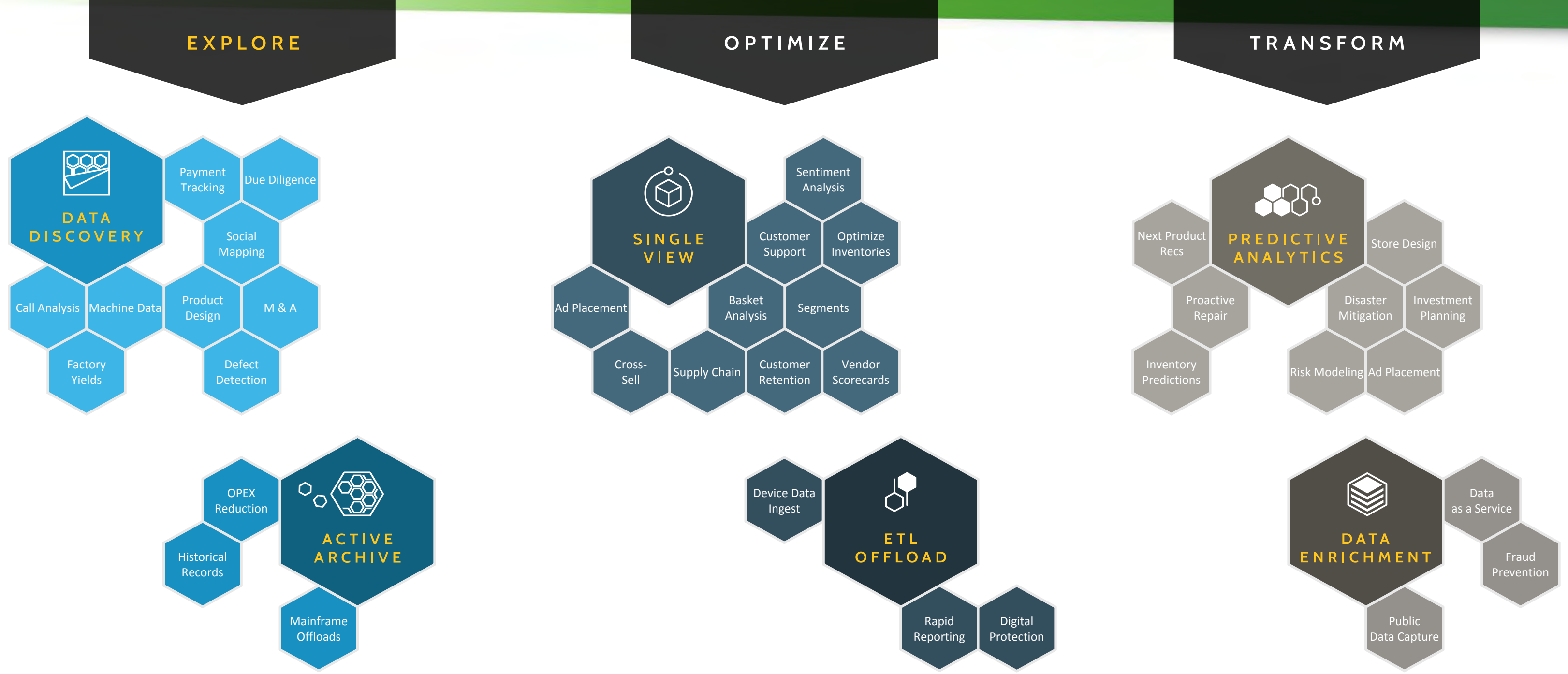


DATA ENRICHMENT

Data as a Service

Fraud Prevention

Public Data Capture



CUSTOMER JOURNEY

Hortonworks® customers leverage our technology to transform their businesses, either by achieving new business objectives or by reducing costs. The journey typically involves both of those goals in combination, across many use cases.

New Analytic Applications for New Types of Data

Financial Services

- New Account Risk Screens
- Fraud Prevention
- Trading Risk
- Maximize Deposit Spread
- Insurance Underwriting
- Accelerate Loan Processing

Retail

- 360° View of the Customer
- Analyze Brand Sentiment
- Localized, Personalized Promotions
- Website Optimization
- Optimal Store Layout

Telecom

- Call Detail Records (CDRs)
- Infrastructure Investment
- Next Product to Buy (NPTB)
- Real-time Bandwidth Allocation
- New Product Development

Manufacturing

- Supplier Consolidation
- Supply Chain and Logistics
- Assembly Line Quality Assurance
- Proactive Maintenance
- Crowdsourced Quality Assurance

Healthcare

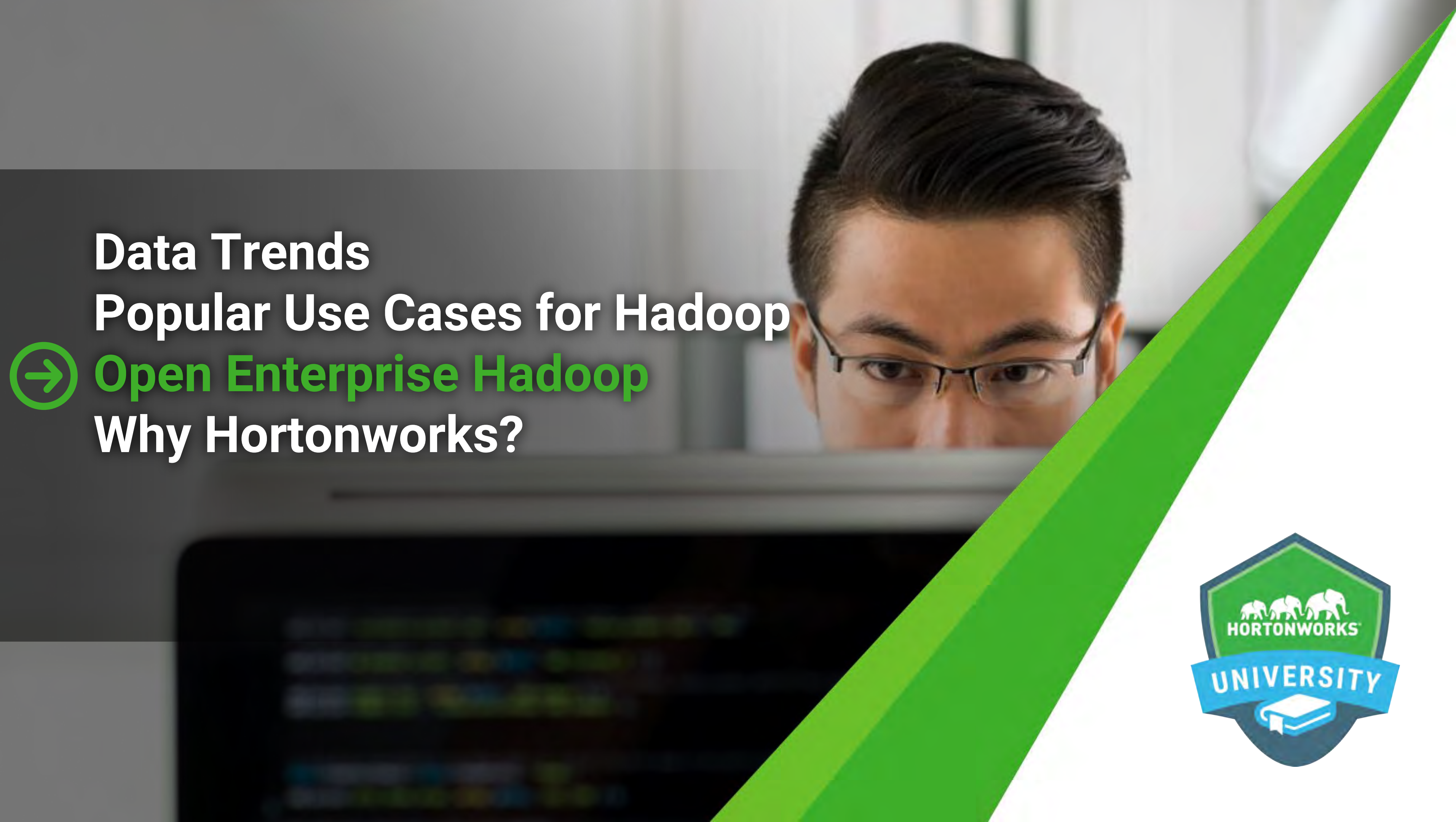
- Genomic data for medical trials
- Monitor patient vitals
- Reduce re-admittance rates
- Store medical research data
- Recruit cohorts for pharmaceutical trials

Utilities, Oil & Gas

- Smart meter stream analysis
- Slow oil well decline curves
- Optimize lease bidding
- Compliance reporting
- Proactive equipment repair
- Seismic image processing

Public Sector

- Analyze public sentiment
- Protect critical networks
- Prevent fraud and waste
- Crowdsource reporting for repairs to infrastructure
- Fulfill open records requests



Data Trends

Popular Use Cases for Hadoop

→ Open Enterprise Hadoop

Why Hortonworks?

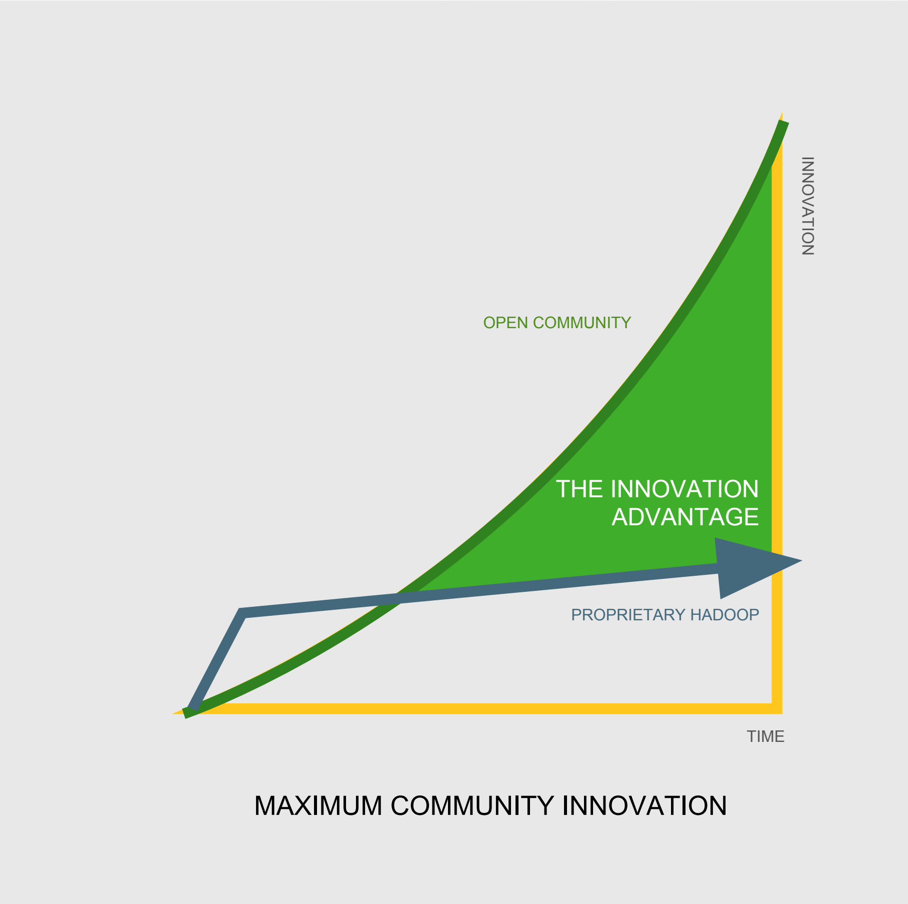
HORTONWORKS UNIVERSITY

Open
Enterprise
Hadoop

- Open
- Central
- Interoperable
- Ready

Open Enterprise Hadoop

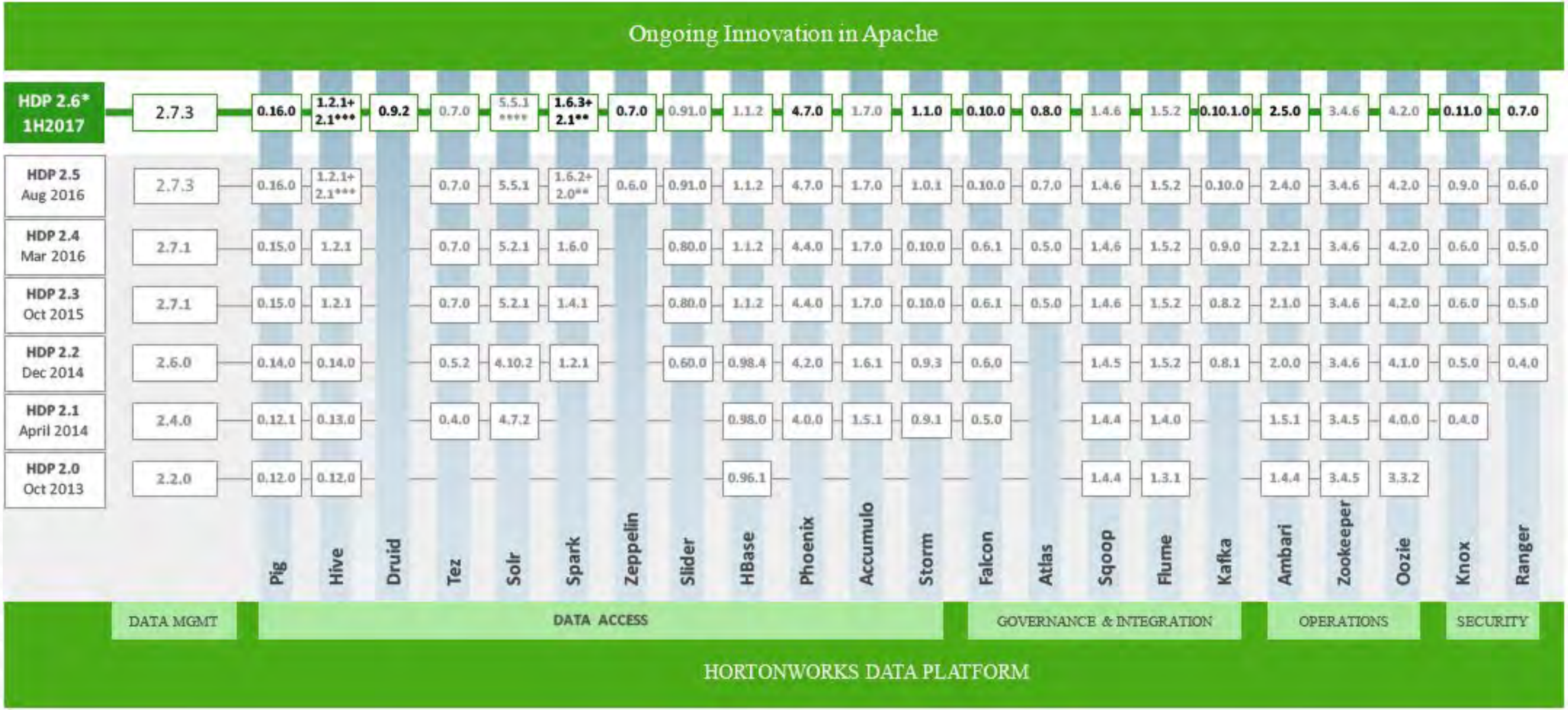
- Open
- Central
- Interoperable
- Ready



Hortonworks Data Platform Is Genuinely Open

- Eliminates Risk
  - of vendor lock-in by delivering 100% Apache open source technology
- Maximizes Community Innovation
  - with hundreds of developers across hundreds of companies
  - **Integrates Seamlessly**
  - through committed co-engineering partnerships with other leading technologies

100% Open Approach = Fastest Path to Innovation



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		

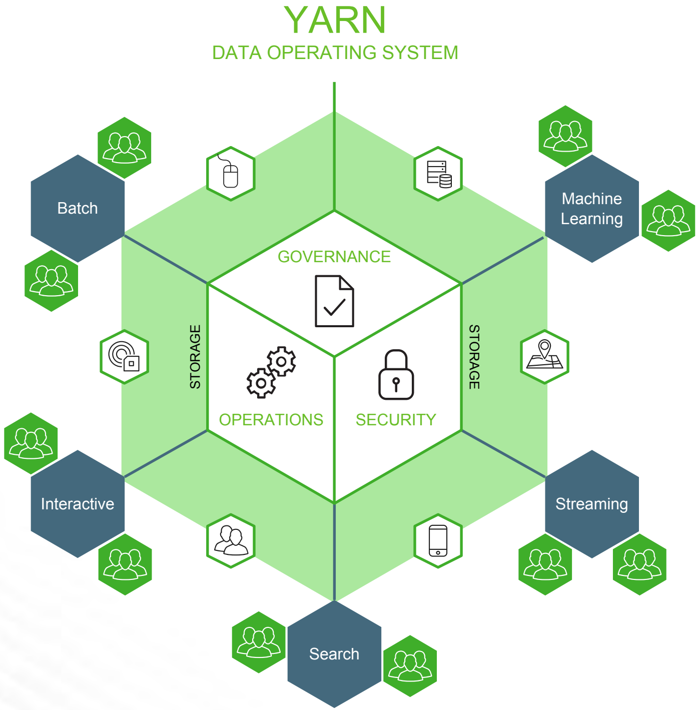
DATA MGMT | DATA ACCESS | GOVERNANCE & INTEGRATION | OPERATIONS | SECURITY

HORTONWORKS DATA PLATFORM

Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Ready

Centralized Platform with YARN-Based Architecture



Centralized Platform

for operations, governance and security

Diverse Applications

run simultaneously on a single cluster

Maximum Data Ingest

including existing and new sources, regardless of raw format

Shared Big Data Assets

across business groups, functions and users

Benefits of the YARN-Based Architecture

		OPEN ENTERPRISE HADOOP 100% YARN-Based Architecture	PROPRIETARY HADOOP Architecture in Silos
Consistent Services	Governance Security Operations	**Confidence with consistent policies for cluster and data management**	Fragmented architecture for the key services **increases complexity and risk**
Resource Efficiency	Management Hardware People	**Shared storage and processing minimizes total cost of ownership**	Redundant clusters mean **more hardware, more data movement,** and **more cost**
Ease of Expansion	Engines Applications Clusters	**Streamlined cluster deployment and also a steady stream of new big data apps to run on YARN**	New applications require additional clusters which **slows deployment and integration**

Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Ready

Offering You the Most Flexibility

ANY DATA

Existing and new datasets

- Click-stream
- Sensor
- Social
- Mobile
- Geo-Location
- Server Log

ANY APPLICATION

Multiple engines for data analysis

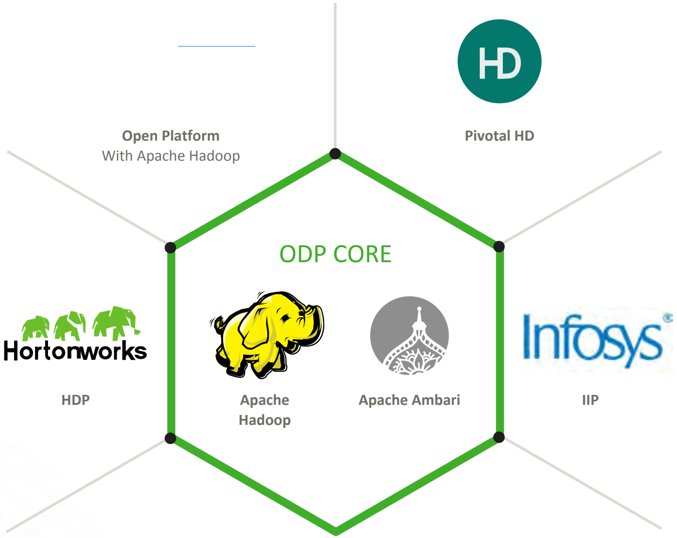
- Batch
- Interactive
- Search
- Streaming
- Machine Learning

ANYWHERE

Complete range of deployment options

- On-Premise
- Cloud
- Linux

Synchronized with Industry Standards



Improves Ecosystem Interoperability

as part of the Open Data Platform (ODP) initiative, founded by Hortonworks

Unlocks Choice

for the customer to use components from multiple vendors integrated with HDP

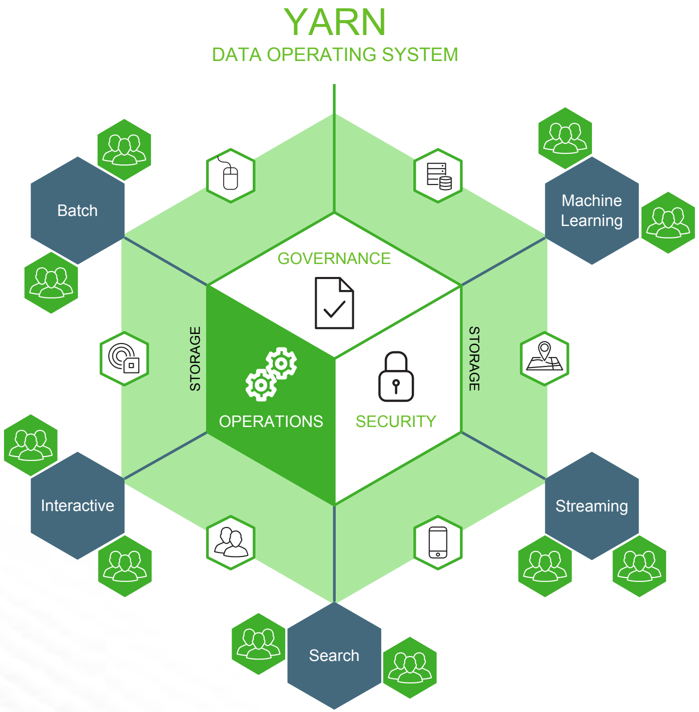
Eliminates Wasteful Guesswork

for the architect who needs to coordinate system versions

Open Enterprise Hadoop

- Open
- Central
- Interoperable
- Ready

Provides Consistent Operations



Centralized

management and monitoring of Hadoop clusters

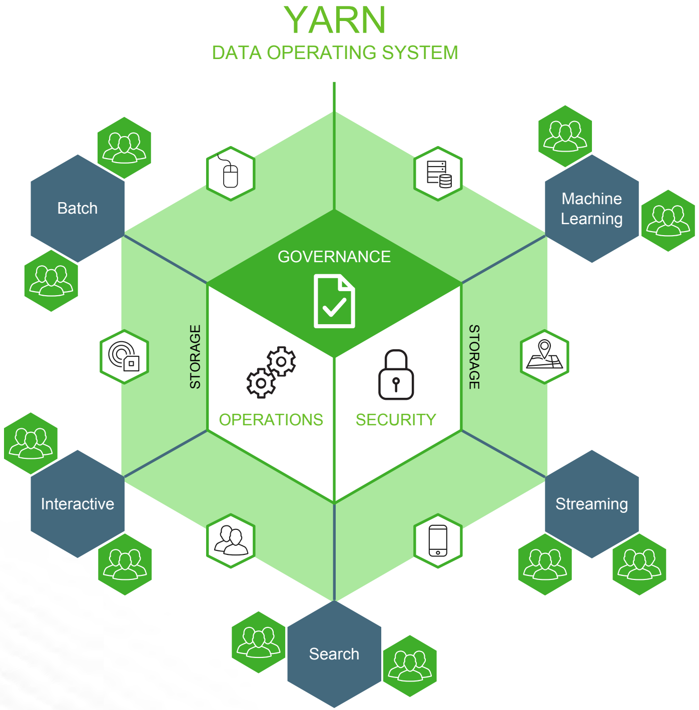
Automated Provisioning

either on-premises or in the cloud with the Cloudbreak API for clusters in minutes

Managed Services

for high availability and consistent lifecycle controls, with dashboards and alerts

Enables Trusted Governance



Data Management

along the entire data lifecycle

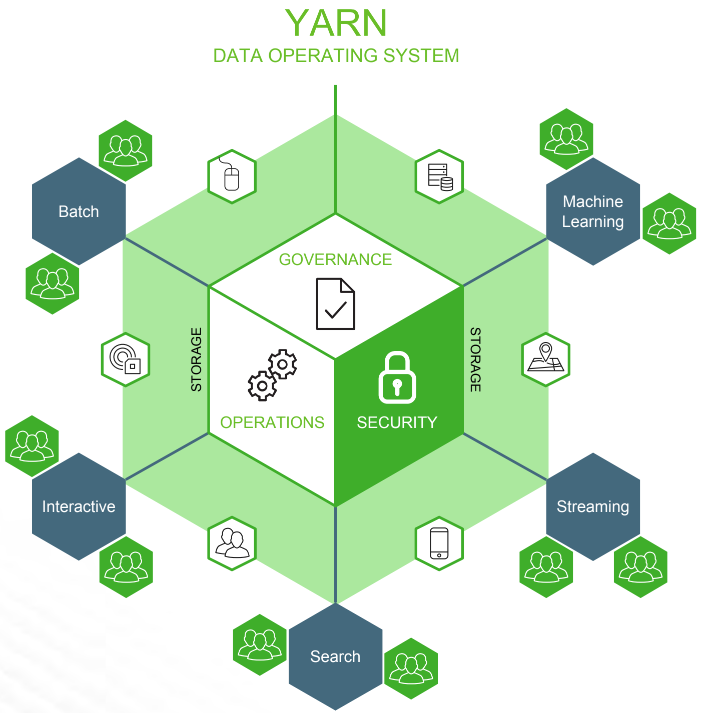
Modeling with Metadata

enables comprehensive data lineage through a hybrid approach

Interoperable Solutions

across the Hadoop ecosystem, through a common metadata store

Ensures Comprehensive Security



Comprehensive Security

through a platform approach

Encrypted Data

at rest and in motion

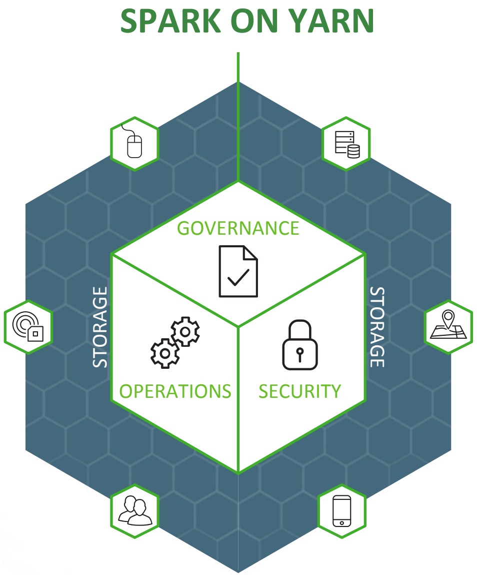
Centralized Administration

of security policies and user authentication

Fine-Grain Authorization

for data access control

Agile Analytics with Enterprise Spark at Scale



Powering Agile Analytics

via data science notebooks and automation for most common analytics (including geospatial and entity resolution)

Seamless Data Access

across as many data types as possible

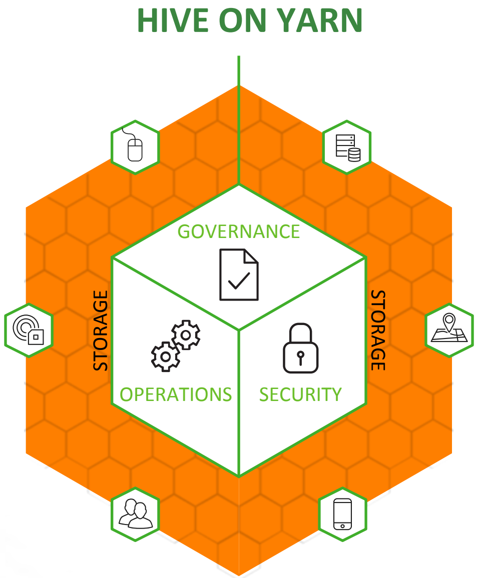
Unmatched Economics

Combining in-memory processing speed with HDP's cost efficiencies at scale

Ready for the Enterprise

with robust security, governance and operations coordinated centrally by Apache Hadoop and YARN

Fast SQL with Apache Hive



Pluggable Architecture

supports Apache Hive, Pivotal HAWQ and other leading SQL engines

Familiar SQL Query Semantics

enable transactions and SQL:2011 Analytics for rich reporting

Unprecedented Speed at Extreme Scale

returns query results in interactive time, even as data sets grow to petabytes

Data Trends

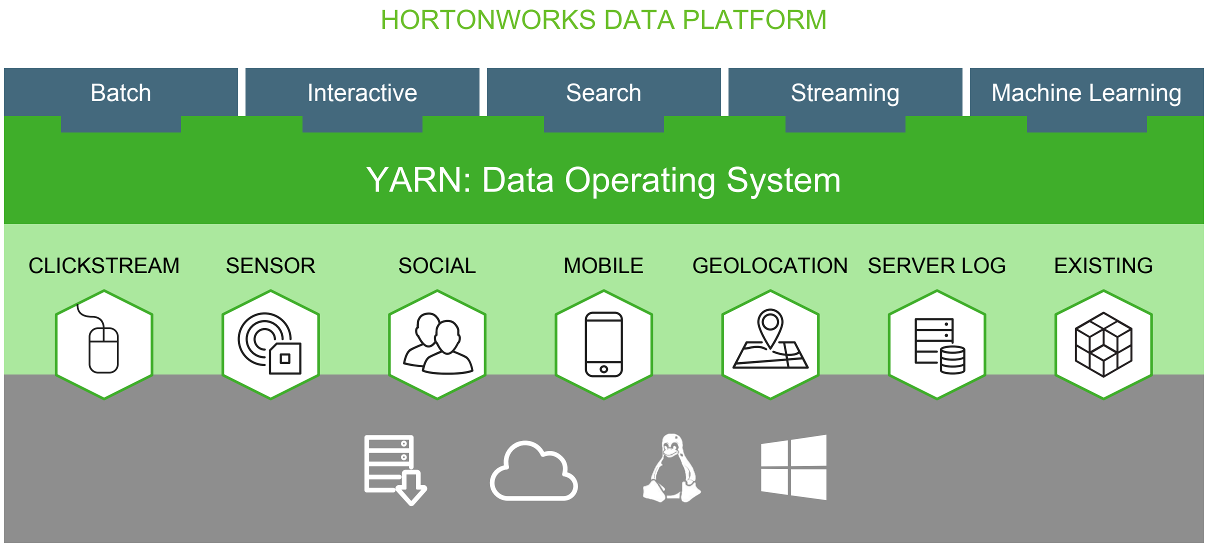
Popular Use Cases for Hadoop

Open Enterprise Hadoop

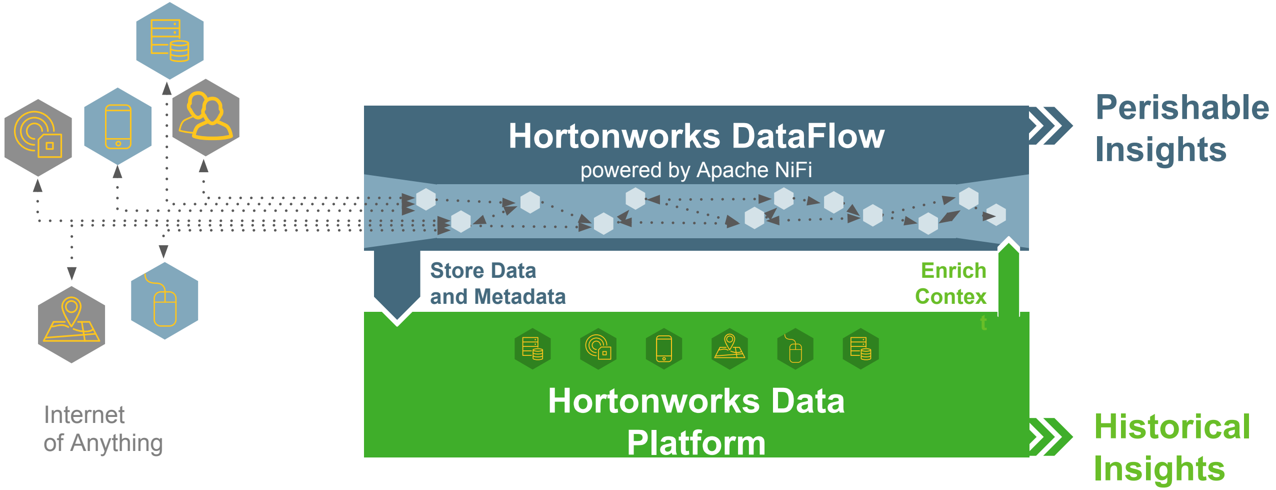
→ Why Hortonworks?

HORTONWORKS UNIVERSITY

Only Hortonworks Delivers Open Enterprise Hadoop



Hortonworks DataFlow Adds to Hadoop Capabilities



Hortonworks DataFlow
powered by Apache NiFi

Perishable Insights

Store Data and Metadata

Enrich Contex
t

Hortonworks Data Platform

Historical Insights

Internet of Anything

Hortonworks DataFlow and Hortonworks Data Platform
deliver the industry's most complete solution for Big Data management

Knowledge Check

Questions

1. List and explain what the “3 V’s” are.

Questions

1. List and explain what the “3 V’s” are.
2. What are the next two data size classifications beyond Petabyte?

Questions

1. List and explain what the “3 V’s” are.
2. What are the next two data size classifications beyond Petabyte?
3. What organization manages Hadoop and its ecosystem of tools and frameworks?

Questions

1. List and explain what the “3 V’s” are.
2. What are the next two data size classifications beyond Petabyte?
3. What organization manages Hadoop and its ecosystem of tools and frameworks?
4. Identify two of the six common use case families.

Questions

1. List and explain what the “3 V’s” are.
2. What are the next two data size classifications beyond Petabyte?
3. What organization manages Hadoop and its ecosystem of tools and frameworks?
4. Identify two of the six common use case families.
5. What one of these use case families is the most widely sought after?

Summary

Summary

- The 3V's of Big Data are driving the adoption of Apache Hadoop (44 ZB by 2020)
- Existing data architectures make data inaccessible, incomplete, irrelevant, and expensive
- Hadoop is a scalable, fault tolerant, open source framework for the distributed storing and processing of large sets of data on commodity hardware
- Six common use case families have emerged
  - Data Discovery
  - Single View
  - Predictive Analytics
  - Active Archive
  - ETL Offload
  - Data Enrichment
- YARN-centralized HDP = Open Enterprise Hadoop