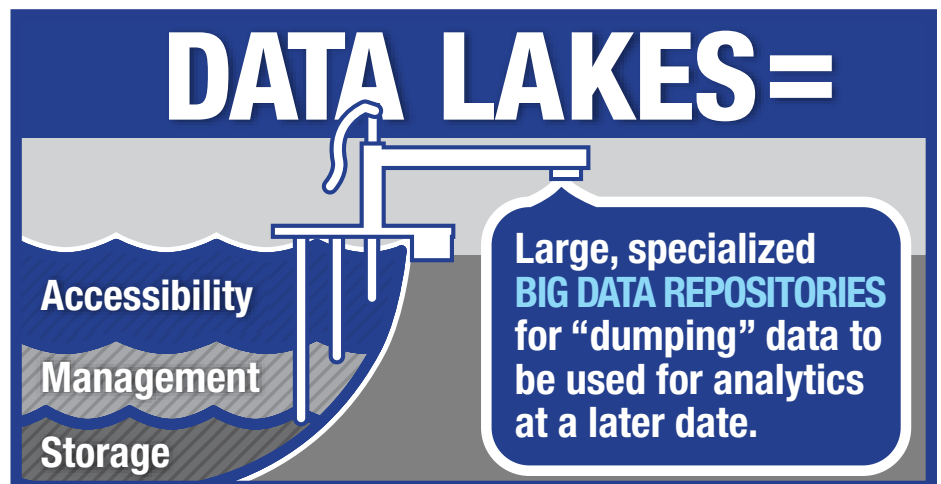# THE JOURNEY TO A DATA LAKE

# 1

## 85% OF DATA GROWTH BY 2020 WILL COME FROM NEW TYPES OF DATA

# THE JOURNEY TO A
# DATA LAKE

ACCORDING TO IDC, AS MUCH AS 85% OF DATA GROWTH BY 2020 WILL COME FROM NEW TYPES OF DATA, ENABLING A TIMELY OPPORTUNITY FOR ENTERPRISES TO DEPLOY HADOOP AND PARTNER WITH INNOVATIVE LEADERS SUCH AS IIS AND HORTONWORKS® TO TAKE A JOURNEY TO THE DATA LAKE.

The IIS and Hortonworks® partnership provides the right expertise to shepherd enterprises on their journey to the data lake, a modern data architecture that addresses many of the shortcomings of today's overwhelmed enterprise data warehouses. In this executive summary, the capabilities of Apache™ Hadoop® as the next-generation, modern enterprise data platform come to light alongside key opportunities. Here are the main takeaways from Hortonworks® ' recent white paper "A Modern Data Architecture with Apache™ Hadoop®: The Journey to a Data Lake."

## DATA LAKES=

Accessibility

Management

Storage

Large, specialized BIG DATA REPOSITORIES for "dumping" data to be used for analytics at a later date.

**iis** TECHNOLOGY
*MISSION POSSIBLE*

Hortonworks

Follow
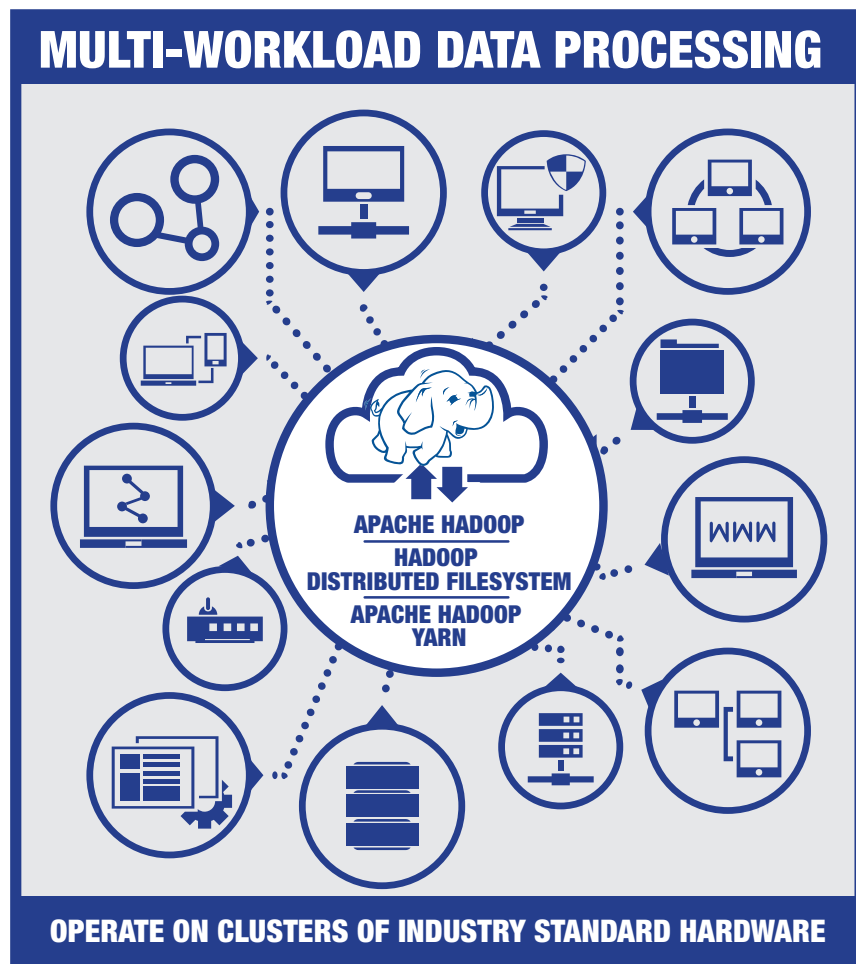
## WHAT IS A DATA LAKE?

Data lakes are quickly becoming the modern data architecture of the next generation of enterprise data platforms, addressing many of the shortcomings of today's enterprise data warehouses. Truly, an enterprise data lake provides new efficiencies and opportunities for data-driven business decisions and innovation.

Data lakes provide a means of leveraging significantly lower cost of storage as well as optimizing workloads to enable other efficiencies. Data lakes allow enterprises to create multi-use and multi-workload data processing engines to perform interactive and real-time processes alongside batch workloads on the same data set.

These capabilities are the result of the maturation of Apache Hadoop as an enterprise-class, Open Source data platform. At its core, Apache™ Hadoop® consists of two key components: Hadoop Distributed Filesystem (based in Java and designed to operate on clusters of industry standard hardware) and Apache Hadoop YARN, which enables multi-workload data processing engines in the data lake. These innovations are the direct result of the data challenges that enterprises must resolve over the next five years.



MULTI-WORKLOAD DATA PROCESSING

APACHE HADOOP
HADOOP DISTRIBUTED FILESYSTEM
APACHE HADOOP YARN

OPERATE ON CLUSTERS OF INDUSTRY STANDARD HARDWARE

Follow

Contact us at www.iisl.com

IIS TECHNOLOGY
MISSION POSSIBLE

Hortonworks

# 3

## THE ERA OF BIG DATA HAS ARRIVED.

### DISRUPTION IN THE DATA

Experts agree that the "era of big data" has arrived. Exponential data growth  is the reality facing enterprises across a broad range of industries,  from retail to manufacturing to life sciences.  For example, in 2012 analysts from IDC estimated the size of the digital universe at 2.8 zettabytes, but by 2020, the digital universe will swell to as many as 40 zettabytes.

## SIZE OF DIGITAL UNIVERSE

**2012:**
**2.8 ZETTABYTES**
**2020:**
**40 ZETTABYTES**

**40 ZB**

**1 ZB=10007BYTES**
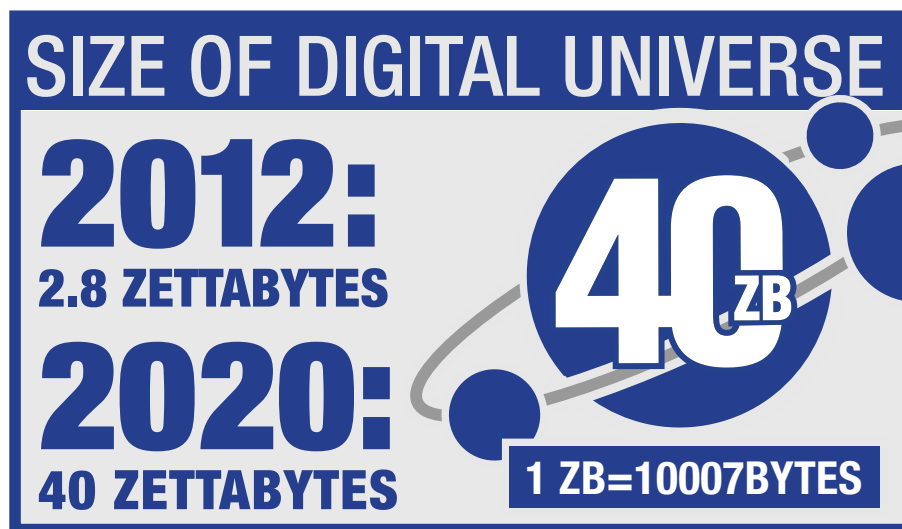
These data, originating from large-scale enterprise applications such as Enterprise Resource & Planning (ERP) systems and Customer Relationship Management (CRM) systems among others, are inundating traditional data warehouses.  This reality has opened a huge opportunity to unleash the value of innovative use cases for Hadoop in the enterprise in partnership with IIS and Hortonworks®.

IIS
TECHNOLOGY
*MISSION POSSIBLE*

Hortonworks

## HADOOP AND A MODERN DATA ARCHITECTURE

As an open source supported technology, Hadoop is a compelling offering to relieve the burden on enterprise data warehouses. With the ability to scale-out at an industry standard on servers, Hadoop truly enables a modern data architecture, which is already in use on several of the largest and most well known web properties on the planet.

Hadoop enables new, innovative opportunities for data analytics. For example, Hadoop uses a schema-on-read capability, which addresses the schema-on-write limitation of traditional data warehouses. Analysts can store raw data and create a schema-upon-analysis, which allows enterprises to address lack of structure in new data and their initial value.

Likewise, Hadoop is capable of accommodating multi-use case, multi-workload data processing engines, which addresses the criticism of Hadoop as primarily a batch-based, single-use case system. On a common data set, analysts can utilize real-time and interactive processes in addition to batch workloads, delivering faster time-to-insight.

## HADOOP ADVANTAGES

**SCHEMA-ON-READ CAPABILITY**

**ADDRESS LACK OF STRUCTURE IN NEW DATA & THEIR INITIAL VALUE**

**ACCOMMODATE MULTI-USE CASE, MULTI-WORKLOAD DATA PROCESSING ENGINES**

**FASTER TIME-TO-INSIGHT**

**OPERATES ON INDUSTRY STANDARD HARDWARE AND DIRECT-ATTACHED STORAGE BY DESIGN**

**IMPROVE STORAGE COSTS**

Hadoop also operates on low-cost, commodity hardware and direct-attached storage by design. This feature allows enterprises to improve storage costs, particularly when compared to legacy technologies such as storage area networks. As companies grow, hardware can scale-out likewise to support the business.
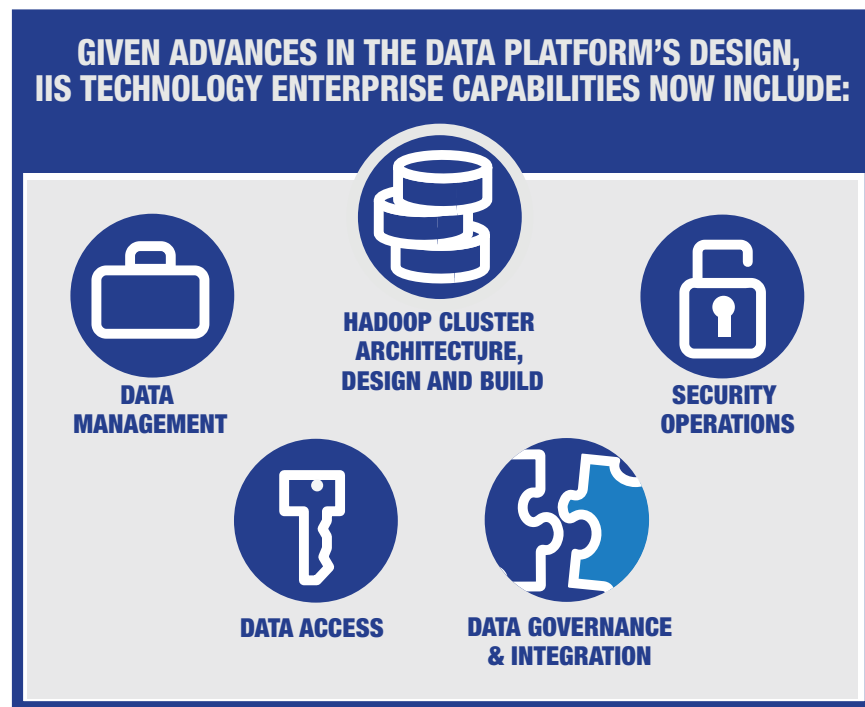
**4**

*HADOOP ENABLES NEW, INNOVATIVE OPPORTUNITIES FOR DATA ANALYTICS.*

**iis**
TECHNOLOGY
**MISSION POSSIBLE**

**Hortonworks**

Follow

# 5

## THE HOT TOPICS IN BIG DATA ARE HADOOP AND NEWSQL-BASED SYSTEMS

### THE BLUEPRINT FOR ENTERPRISE HADOOP

*IIS AND HORTONWORKS® CAN PROVIDE BUSINESSES WITH THE RIGHT BLUEPRINT FOR HADOOP IN THE ENTERPRISE.*

**GIVEN ADVANCES IN THE DATA PLATFORM'S DESIGN, IIS TECHNOLOGY ENTERPRISE CAPABILITIES NOW INCLUDE:**

DATA MANAGEMENT

HADOOP CLUSTER ARCHITECTURE, DESIGN AND BUILD

SECURITY OPERATIONS

DATA ACCESS

DATA GOVERNANCE & INTEGRATION

In conjunction with the Open Source community and major technology partners, such as Microsoft, Rackspace and HP, Hadoop has matured substantially as an enterprise-class data platform. Established and new vendors have joined the Hadoop movement and made contributions through the Apache Software Foundation.

Hortonworks® Data Platform (HDP™) is the only completely open Hadoop data platform available. HDP™ delivers Enterprise Apache™ Hadoop®, deeply integrated with existing systems to create a highly efficient, highly scalable way to manage all your enterprise data.

A thriving ecosystem of vendors ensures that Hadoop will continue to improve upon its core components. As many of Hadoop's key vendors are already established within the data warehouse space, enterprise Hadoop will only benefit from this growing ecosystem. As leaders in Hadoop technology, with strong partnerships and support, IIS and Hortonworks® make significant contributions to the community and the big data ecosystem.

**IIS** TECHNOLOGY
**MISSION POSSIBLE**

**Hortonworks**

Follow

## TOWARD A DATA LAKE

Data lakes are a modern data architecture enabled by Hadoop's enterprise capabilities. Hadoop is currently in use on some of the largest clusters on the planet. The use cases for Hadoop in the enterprise abound. For example, a retailer could deploy Hadoop to analyze clickstream, geolocation and social media data to gain a true 360-degree view of customers, dynamically and faster than ever before.



**HADOOP FOR RETAIL**

DYNAMIC & FAST 360-DEGREE VIEW OF CUSTOMERS
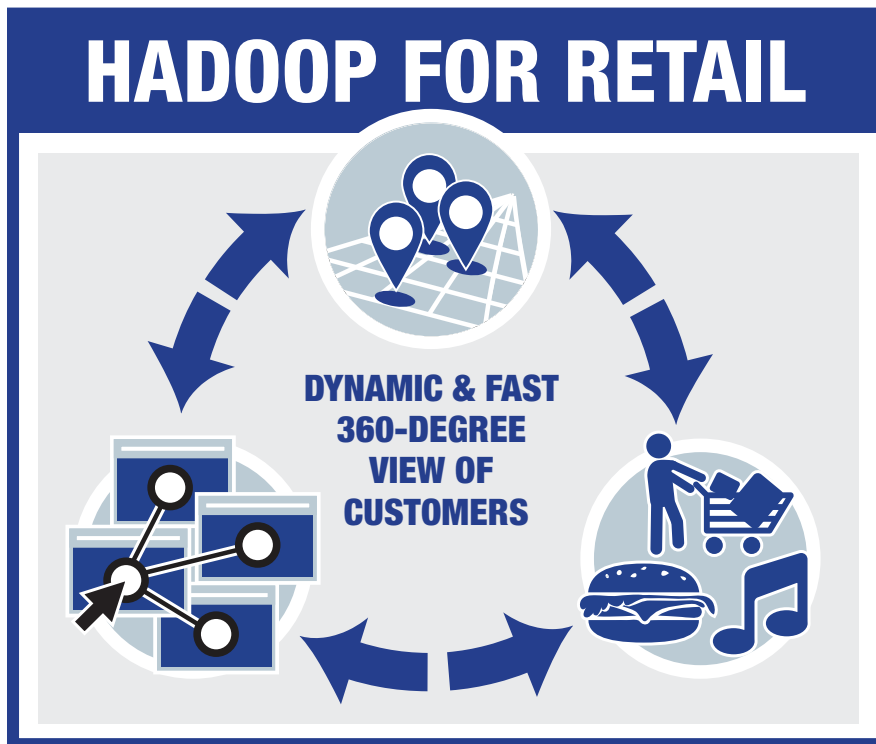
Data lakes allow enterprises to collect and analyze all data regardless of schema. There is no such thing as "junk data" any longer. Enterprises can access and glean insights from data in innovative ways with a data lake and Hadoop. In the end, data lakes create the opportunity to deliver maximum insights at maximum scale at the lowest possible cost.

**6**

*DATA LAKES ALLOW ENTERPRISES TO COLLECT AND ANALYZE ALL DATA REGARDLESS OF SCHEMA.*

**iis** TECHNOLOGY
*MISSION POSSIBLE*

**Hortonworks**

Follow

# 7

## THE VALUE OF HORTONWORKS' AND IIS' PARTNERSHIP

Through their partnership, IIS and Hortonworks® seek to help enterprises on their journey to the data lake. As the largest contributor to the Hadoop community, Hortonworks® stands uniquely poised to foster the next generation of use cases for Hadoop. Enterprises that choose to join the forefront of innovation with IIS and Hortonworks® will find much value in the solutions these two industry leaders provide.

To solve today's data challenges, enterprises need to deploy a modern data architecture. Hadoop has quickly taken the lead as the go-to data platform to complement enterprise data warehouses today and in the future. IIS now offer a strong portfolio of services with Hadoop and HDP™ to guide enterprises to the data lake effectively and efficiently.

*FOR MORE INFORMATION ON HOW IIS AND HORTONWORKS CAN HELP ENTERPRISES ON THEIR JOURNEY TO THE DATA LAKE, VISIT WWW.IISL.COM TO CONTACT A REPRESENTATIVE.*

**IIS**
TECHNOLOGY
*MISSION POSSIBLE*

**Hortonworks**

**IIS**
TECHNOLOGY
*MISSION POSSIBLE*

**Hortonworks**