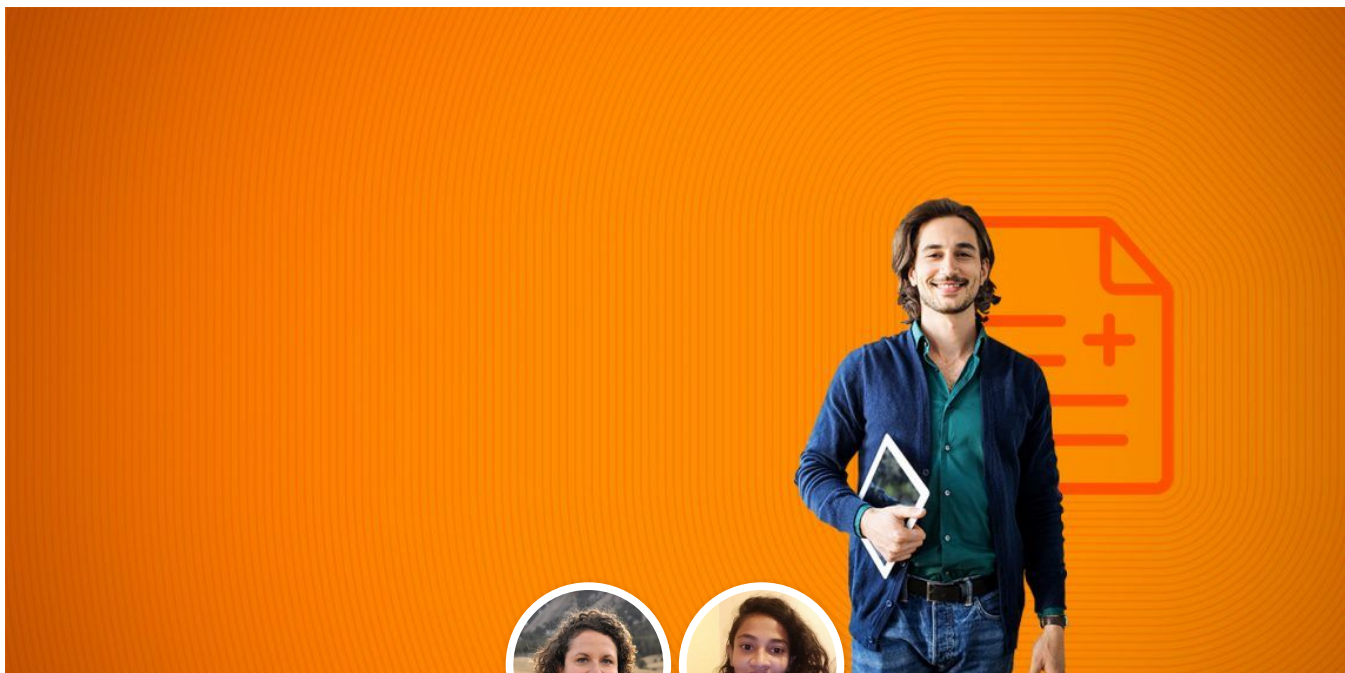| AIOps | BMC Beat | Cloud | DevOps | Innovation | ITSM | Mainframe | Workload Automation |

More ⌄

Machine Learning & Big Data Blog

Hadoop vs Kubernetes: Will K8s & Cloud Native End Hadoop?

June 18, 2021

6 minute read

Chrissy Kidd, Shanika Wickramasinghe

Apache Hadoop is one of the leading solutions for distributed data analytics and data storage. However, with the introduction of other distributed computing solutions directly aimed at data analytics and general computing needs, Hadoop's usefulness has been called into question.

There are many debates on the internet: is Hadoop still relevant? Or, is it dead altogether?

In reality, Apache Hadoop is not dead, and many organizations are still using it as a robust dat ⌃

AIOps　　BMC Beat　　Cloud　　DevOps　　Innovation　　ITSM　　Mainframe　　Workload Automation

More ⌄　　🔍

Hadoop Guide ⌃

Beginning Kubernetes: Knowledge & Tutorials for Getting Started

n Hadoop reached its peak popularity from 2014 to 2017. searches for Hadoop. However, this alone is not a good the current landscape. After all, Hadoop can be integrated olete analytics solution.

Note
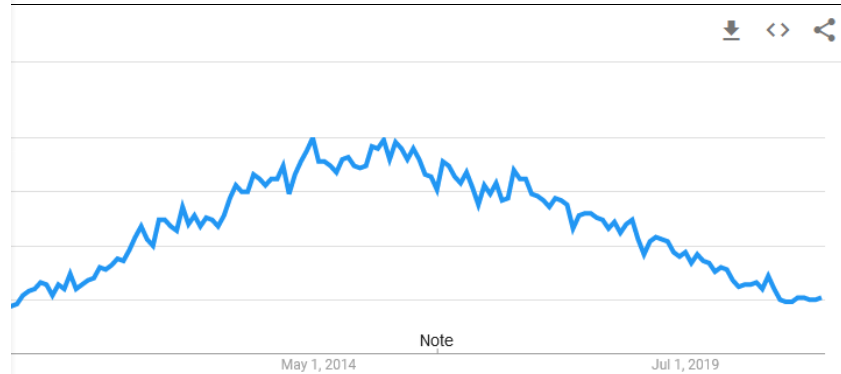May 1, 2014　　　　　　Jul 1, 2019

In this article, we will learn more about Hadoop, its usability, and whether it will be replaced by rapidly evolving technologies like Kubernetes and Cloud-Native development.

Download e-book ›

(This article is part of our Hadoop Guide. Use the right-hand menu to navigate.)

What is Hadoop?

Hadoop is an open-source framework that is used to store and process massive datasets efficiently. It is a reliable and scalable distributed computing platform that can be used on commodity hardware.

Hadoop distributes its data storage and analytics workloads across multiple nodes (computers) to handle the work parallelly. This leads to faster, highly efficient, and low-cost data analytics capabilities.

Hadoop modules

Hadoop consists of four main modules that power its functionality:

HDFS. Hadoop Distributed File System is a file system that can run on low-end hardware while providing better throughput than traditional file systems. Additionally, it has built-in fault tolerance and the ability to handle large datasets.

and resource management of the cluster.

MapReduce. MapReduce is a big data processing engine that supports the parallel More processing of large data sets. It is the default processing engine available on Hadoop. Hadoop Guide Currently, however, Hadoop also provides support for other engines such as Apache Tez

Common provides a common set of libraries that can be used odules.

behind the popularity of Apache Hadoop.

distributed computing model allows it to handle limitless

Data safety. Hadoop automatically creates and manages data backups. So, you can simply recover your data from a backup in case of a failure.

Cost. Hadoop's ability to run on commodity hardware enables organizations to easily deploy a data analytics platform using it. It also eliminates the need for expensive and sp

Availability. Hadoop is designed to handle failures at the application layer—which means it provides high availability without relying on hardware.

With its flexibility and scalability, Hadoop quickly gained the favor of both individual data engineers/analysts and corporations. This flexibility extends to types of data Hadoop can collect:

Structured and unstructured
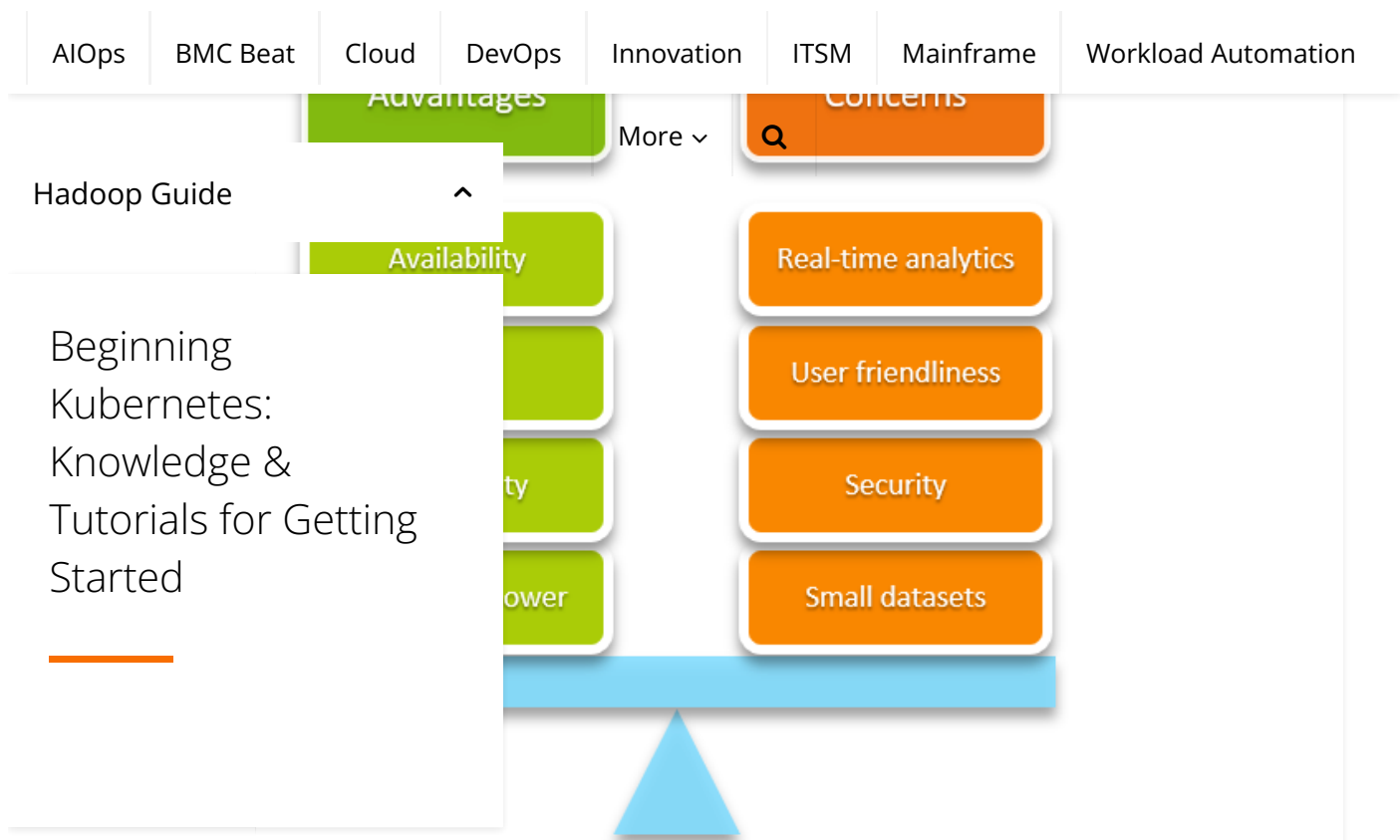
Through different inputs like social media, streamed data, internal collections, etc.

Then Hadoop checks all these data sets and determines the usefulness of each data set. All this is done *without* having to go through the process of converting data into a single format.

Another feature that elevates Hadoop is its storage capability.

Once a large data set is accumulated, and the required data is extracted, we can simply store the unprocessed data with Hadoop endlessly. This enables users to reference older data easily, and the storage costs are also minimal since Hadoop is running on commodity hardware.

bmc

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

Advantages

More ⌄   🔍

Hadoop Guide   ^

Availability

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

Concerns

Real-time analytics

User friendliness

Security

Small datasets

Drawbacks of Hadoop

Download e-book ›

Apache _____ ___ ___ed prominence thanks to all the above features.

However, as technology advances, new options have emerged, challenging Hadoop and even surpassing it in certain aspects. This, along with the inherent limitations of Hadoop, means it has indeed lost its market lead.

So, what are some drawbacks of Hadoop?

Inefficient for small data sets

Hadoop is designed for processing big data composed of huge data sets. It is very inefficient when processing smaller data sets. Hadoop is not suited and cost-prohibitive when it comes to quick analytics of smaller data sets.

Another reason: Although Hadoop can combine, process, and transform data, it does not provide an easy way to output the necessary data. This limits the available options for business intelligence teams for visualizing and reporting on the processed data sets.

Security concerns

⌃

decryption at the storage or network levels. Only Kerberos authentication is officially supported by Hadoop, which is a technology that is difficult to maintain by itself.

Hadoop Guide   ^

... each Hadoop configuration, users need to manually enable security options or use third-party ...

Beginning Kubernetes: Knowledge & Tutorials for Getting Started

—————

...ess

...e of the leading programming languages with a large ...t the best language for data analytics, and it can be complex

...nfigurations and usage—the user must have thorough ... to properly use and debug the cluster.

Not suitable for real-time analytics

Hadoop is designed with excellent support for batch processing. However, with its limitations in process[ing ...] *and* not providing native support for real-time analytics, Hadoop is ill-suited [...quick real-time]analytics.

Download e-book ›

Hadoop alternatives

So, what other options to Hadoop are available? While there is no single solution to replace Hadoop outright, there are newer technologies that can reduce or eliminate the need for Hadoop.

Apache Spark

Apache Spark is one solution, provided by the Apache team itself, to replace MapReduce, Hadoop's default data processing engine. Spark is the new data processing engine developed to address the limitations of MapReduce.

Apache claims that Spark is nearly 100 times faster than MapReduce and supports in-memory calculations. Moreover, it supports real-time processing by creating micro-batches of data and processing them.

The support of Spark for modern languages enables you to interact using your preferred programming languages. Spark offers excellent support for data analytics using languages suc ^

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

Scala   More ⌄   🔍

Hadoop Guide   ⌃

Spark SQL

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

⎯⎯⎯⎯

e Flink. Flink is another processing engine with the same
higher performance in some workloads as it is designed to
nbounded and bounded data streams.

& cloud-native replace Hadoop?

Even with newer and faster data process engines, Hadoop still limits users to its tools and technologies like HDFS and YARN with Java-based tools. But, what if you need to integrate other tools and

Download e-book ›

e best for your specfic data storage and analytics needs?

The solution is using Kubernetes as the orchestration engine to manage your cluster.

With the ever-growing popularity of containerized cloud-native applications, Kubernetes has become the leading orchestration platform to manage any containerized application. It offers features such as:

Convenient management

Networking

Scaling

High availability

(Explore our comprehensive Kubernetes Guide.)

Consider this scenario: you want to move to cheap cloud storage options like Amazon S3 buckets and managed data warehouses like Amazon Redshift, Google BigQuery, Panoply. This is not possible with Hadoop.

Kubernetes, meanwhile, can easily plug them into Kubernetes clusters to be accessed by the   ⌃

responsibilities as cloud providers manage all the day-to-day maintenance and availability of data

Hadoop Guide   ⌃

Having the storage sorted, Kubernetes can host different services such as:

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

e Spark, Presto, Flink)

yter, NLTK, TensorFlow, PyTorch, MATLAB)

netes cluster

ny tools, frameworks, or programming languages you're
's most suitable for your use case—you're no longer limited

ork together.)

Portability of Kubernetes

Another factor that uplifts Kubernetes is its portability. Kubernetes can be easily configured to be distr[Download e-book ›   ]ocations and run on multiple cloud environments. With containerized applications, users can easily move between development and production environments to facilitate data analytics in any location without major modifications.

By combining Kubernetes with rapid DevOps and CI/CD pipelines, developers can easily create, test, and deploy data analytics, ML, and AI applications virtually anywhere.

Support of Kubernetes for Serverless Computing

Kubernetes has further eliminated the need to manage infrastructure separately with the support for serverless computing. Serverless computing is a rising technology where the cloud platform automatically manages and scales the hardware resources according to the needs of the application.

Some container-native, open-source, and function-as-a-service computing platforms like fn, Apache OpenWhisk, and nuclio can be easily integrated with Kubernetes to run serverless applications—eliminating the need for technologies like Hadoop.

Some frameworks, like nuclio, are specifically aimed at automating data science pipelines with serverless functions.

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

for managing any big data workloads.

More ⌄   🔍

Hadoop Guide   ^

Hadoop handles large data sets cheaply

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

s also designed to address a specific need—handling large
hardware.

s have given rise to new requirements and use cases.
ologies, like Kubernetes and serverless computing, offer
tions.

u to identify and utilize the correct technology stack for your

Related reading

[BMC Machine Learning & Big Data Blog](#)

[Data Architecture Explained: Components, Standards & Changing Architectures](#)

3   ient Data Pipelines
        Download e-book ›

[Snowflake 101: Intro to the Snowflake Data Cloud](#)

[Big Data: A Big Introduction](#)

[Data Ethics for Companies](#)

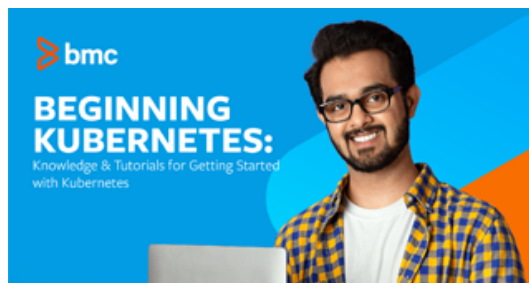Beginning Kubernetes: Knowledge & Tutorials for

More ∨   🔍

Hadoop Guide   ∧

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

e take a deep

ng platform

These postings are my own and do not necessarily represent BMC's position, strategies, or opinion.

See an error or have a suggestion? Please let us know by emailing blogs@bmc.com.

#Hadoo

Download e-book ›

[Facebook]   [Twitter]   [LinkedIn]

BMC Brings the A-Game

BMC works with 86% of the Forbes Global 50 and customers and partners around the world to create their future. With our history of innovation, industry-leading automation, operations, and service management solutions, combined with unmatched flexibility, we help organizations free up time and space to become an Autonomous Digital Enterprise that conquers the opportunities ahead.

Learn more about BMC ›

AIOps	BMC Beat	Cloud	DevOps	Innovation	ITSM	Mainframe	Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

Download e-book ›

Spark's Machine Learning Pipeline: An Introduction

What Is a Neural Network? An Introduction with

How to Add Subplots in Matplotlib

What Is a Canonical Data Model? CDMs Explained

Data Storage Explained: Data Lake vs Warehouse vs
Database

The Data Visualization Beginner's Guide

You may also like

About the author

Chrissy Kidd

Chrissy Kidd is a writer and editor who makes sense of theories and new   ⌃

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

you can reach her on LinkedIn or at ~~emesbykidal.com~~.

More ∨   🔍

Hadoop Guide   ∧

Beginning Kubernetes: Knowledge & Tutorials for Getting Started

in

About the author

kramasinghe

nasinghe is a software engineer by profession and a rmation Technology. Her specialties are Web and Mobile nanika considers writing the best medium to learn and edge. She is passionate about everything she does, loves to travel, and enjoys nature whenever she takes a break from her busy work schedule. You can connect with her on LinkedIn.

in

Vie   Download e-book ›

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

―――――

Download e-book ›

⌃

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

———

Download e-book ›

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

Download e-book ›

⌃

Hadoop Guide   ⌃

Beginning Kubernetes: Knowledge & Tutorials for Getting Started

———

Download e-book ›

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ^

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

_____

Download e-book ›

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

———

Download e-book ›

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

———

Download e-book ›

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

⌃

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

Download e-book ›

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

Download e-book ›

AIOps | BMC Beat | Cloud | DevOps | Innovation | ITSM | Mainframe | Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

Download e-book ›

AIOps | BMC Beat | Cloud | DevOps | Innovation | ITSM | Mainframe | Workload Automation

⌃

AIOps   BMC Beat   Cloud   DevOps   Innovation   ITSM   Mainframe   Workload Automation

More ⌄   🔍

Hadoop Guide   ⌃

Beginning
Kubernetes:
Knowledge &
Tutorials for Getting
Started

Trust Center

Accessibility Tools

Cookie Consent

Download e-book ›