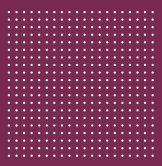
CANONICAL



WHITE PAPER

Ubuntu and Hadoop: the perfect match

February 2012

Executive introduction

In many fields of IT, there are always stand-out technologies. This is definitely the case in the Big Data space, where Hadoop is leading the way.

While a range of individual data applications can help organisations gain insight from unstructured data, Hadoop goes much further. It provides an entire software ecosystem for Big Data analysis – co-ordinating any number of physical or virtual machines to deliver business insight.

IBM have positioned three elements for Big Data (the 'three Vs'). These are data Variety, data Volume, and processing Velocity. Hadoop helps organisations overcome all of these challenges. For the first time, it's possible to analyse massive, structurally diverse datasets quickly and cost effectively – lowering the Big Data bar for businesses of all sizes, across all sectors.

Contents

Ubuntu and Hadoop: the perfect match

Ubuntu: the ideal foundation for Hadoop

The true value of Hadoop is that it can co-ordinate distributed infrastructure elements to achieve a common end. However, without the right technology to underpin it, distributed Hadoop storage and processing environments can become complex to deploy and manage, and expensive to license.

Because it is based on open industry standards, optimised for the cloud, and able to scale with no increments in licensing costs, Ubuntu is the ideal foundation for Hadoop clusters. In this paper, we explain exactly why Ubuntu is becoming the dominant platform for Hadoop, and showcase how one organisation, Music Metric, has used the combined technologies to achieve competitive advantage.

Hadoop: the market need

For years, web giants such as Facebook and Google have been analysing unstructured data to gain insight into user preferences, exploit commercial opportunities, and provide a more relevant web experience. However, organisations with fewer IT resources have lacked the tools and processing power to maximise the business value of their own, rapidly growing datasets – until now...

With the advent of commodity hardware and public cloud infrastructure, more organisations than ever have access to super-computing resources. As a result, Big Data analysis is becoming possible for more and more organisations in multiple sectors.

Hadoop is playing a pivotal role in this Big Data revolution.

Critically, it can deliver business value across a range of business applications and industries – anywhere that there's a requirement to store, process, and analyse large volumes of data. It can be used, for example, to streamline and automate digital marketing processes, detect fraud in the financial services industry, or assess customer preferences based on data from social networks.

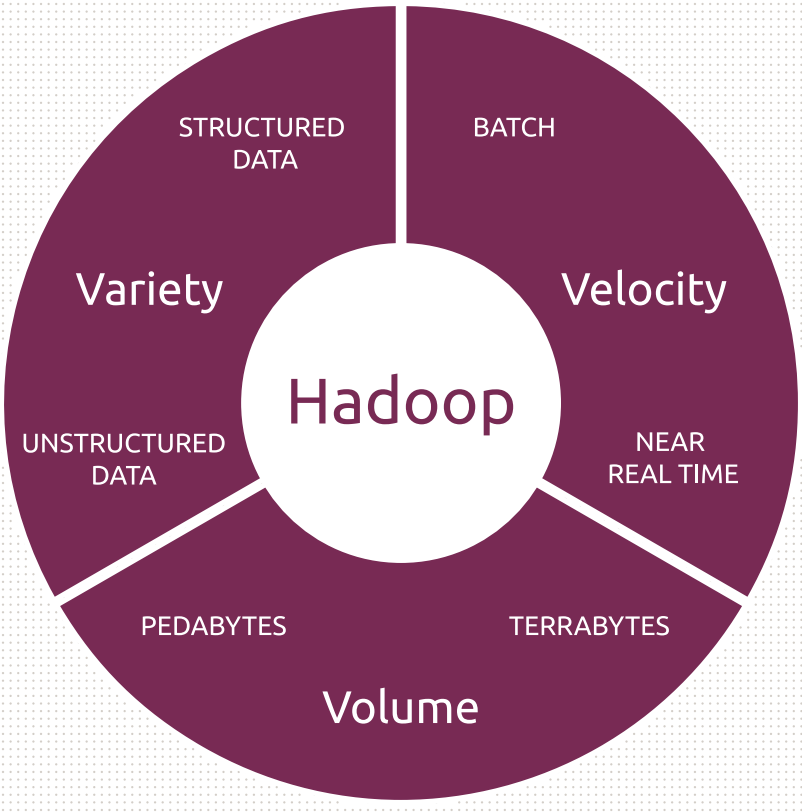
The broad range of business insight enabled by Hadoop has ensured rapid adoption across all industries. Already, the world's leading companies use it – including major online companies, banks, utilities companies and many more.

How does it work?

By enabling organisations to co-ordinate distributed servers to work on massive analysis tasks, Hadoop is playing a significant role in the Big Data revolution. Critically, it helps organisations:

- **Analyse a variety of data types**
  That includes a wide variety of data, parameters and fields. Data may come from extremely diverse sources, from customer databases to logs generated by telephony routers or reports from medical equipment.
- **Process huge volumes of data**
  Hadoop can potentially process unlimited quantities of unstructured data, depending on the availability of servers and other infrastructure resources. It is not uncommon for Hadoop clusters to run on more than a hundred physical or virtual machines.
- **Achieve results quickly**
  Because it scales horizontally across distributed infrastructure, Hadoop is capable of processing data much more quickly than competing technologies. This helps organisations generate specific results from their data faster – such as customer preferences by demographic, geography or other criteria.

IBM has catagorised data as exhibiting up to 3 characteristics:
Volume, Variety, & Velocity:



What is Hadoop?

Hadoop is a software framework that supports data intensive distributed applications. It consists of several components:

- **MapReduce** – an environment for authoring applications that process vast volumes of structured and unstructured data quickly and reliably across large clusters of virtual or physical machines
- **Hadoop Distributed File System (HDFS)** – allows large volumes of data to be stored and rapidly accessed across literally hundreds of nodes
- **Hive** – for ad-hoc queries of datasets stored in HDFS
- **HBase** – a NoSQL environment that enables instant read and write access to Big Data for presentation to end users

Why Ubuntu for Hadoop?

Ubuntu offers a number of distinct advantages for organisations planning to deploy a Hadoop cluster. These include:

1. SUPPORT FOR THE LATEST HADOOP FEATURES

Hadoop is an incredibly fast-moving technology, and new features are being added every day. To make sure organisations have access to all the latest Hadoop tools, their underlying systems – including the OS – must be up to date.

Because Ubuntu follows a regular, six-monthly release cycle, all the latest Hadoop technologies are available to the platform. By contrast, many proprietary systems and commercial Linux-based systems are only upgraded every two or three years (Microsoft, Red Hat and others), creating the possibility that some of the latest Hadoop developments may not be available on those platforms.

2. SCALABILITY WITHOUT LIMITS

Proprietary and commercial open-source systems all follow the same commercial model: use more, pay more. This is an extremely impractical model for Hadoop deployments, which typically require significant computing power, storage and scalability – distributed across physical and virtual machines.

Ubuntu Server offers a cost-effective way of deploying distributed Hadoop clusters, with no need for costly, per-machine OS licenses. Because our technology is free to download and deploy, you can deploy 50, 100 or 500 additional Hadoop nodes in your datacentre or in the cloud to support your Big Data activities without increasing software costs.

We'd recommend you contract enterprise support from Canonical, the company behind Ubuntu, to protect your mission-critical Big Data operations, which obviously has a cost implication. However, downloading and deploying Ubuntu Server is, and will always be, free.

3. NATIVE SUPPORT FOR THE CLOUD

Because Hadoop analyses typically demand intensive computing resources for short periods of time, many organisations decide to deploy it on public clouds. This provides the flexibility to scale resources up and down on demand, and reduces in-house infrastructure requirements and costs.

Ubuntu's native support for private and public clouds makes it an ideal platform for organisations deploying Hadoop. With best-of-breed cloud infrastructure built into Ubuntu Server, it is possible to achieve the required levels of computing elasticity, with real-time resource provisioning and scaling.

4. RAPID DEPLOYMENT

Organisations can deploy Hadoop faster on Ubuntu. This is thanks to a range of innovative tools that have been developed by Canonical – the company behind the Ubuntu project.

Among these, Juju is especially significant. It can be used to deploy Hadoop across a distributed infrastructure in minutes using a simple, pre-written 'charm' that can be downloaded online. Once the service is configured, a systems administrator decides how many back-end databases, web servers and application servers are needed to support it, using easy to use tools to interconnect and deploy them.

To experience the power of Juju first hand, visit: juju.ubuntu.com

5. OUT-OF-THE-BOX HARDWARE SUPPORT

Hadoop, like all Big Data applications, is resource intensive. To reduce the cost of deployment, it needs to run on low-cost, commoditised hardware in-house, or in the cloud.

Ubuntu fulfils this need based on out-of-the-box hardware compatibility. The OS is certified for more than 220 servers and recent x86 CPU devices from all major manufacturers. Since the release of Ubuntu 11.10 in October 2011, Ubuntu Server will also be available to run on ARM-based servers, providing additional power consumption and space-saving options for Hadoop deployments.

Critically, Ubuntu also supports all hardware compatible with Open Compute Project (OCP) standards, which enable open-source-based software to be deployed seamlessly on hardware devices direct from participating manufacturers.

A full list of hardware that is optimised for Ubuntu Server is available at: ubuntu.com/certification

6. RELATIONSHIPS WITH HADOOP DEVELOPERS

Canonical has developed relationships with the leading Hadoop development and distribution companies such as Hortonworks, MapR and Cloudera. In addition, Ubuntu developers at Canonical work with the upstream Apache Hadoop community to ensure that the very latest Hadoop technologies are available to Ubuntu users.

Customer showcase: Music Metric

Music Metric, an innovative London technology company, provides unique insight for music industry players and artists. It does this by analysing vast quantities of online data from music download sites, social networks, and blogs. To collect and process this information on an hourly basis, the organisation operates a large, three-tiered IT infrastructure. This is powered from end-to-end by Ubuntu 10.04 LTS, from cloud-based data collection systems, to the company's Hadoop Big Data analysis cluster and data presentation tools.

With a single OS image supporting Music Metric's entire IT stack, development and routine administration is far simpler. What's more, Ubuntu's long-term support ensures that the company's critical systems are totally stable and constantly available.

Jameel Sayed, CTO of Music Metric, says: "Hadoop on Ubuntu enables us to co-ordinate lots of machines to work together processing and aggregating massive volumes of data. We feed in the raw information at one end and get lots of different measures of artist popularity out the other."

In summary

Hadoop is shaping the future of Big Data, giving businesses of all types and sizes new insight into operational performance, assets, customer preferences and more. Among the Big Data technologies, it is by far the fastest growing – and Hortonworks predicts that nearly half the world's data will be processed and analysed in Hadoop environments within the next five years.[1]

To ensure the success of Hadoop deployments, however, organisations must choose the right underlying technologies. It's important, for example, that the Hadoop file systems and databases can scale across distributed infrastructure – whether in the datacentre or in the cloud – and that software licensing costs are never a barrier to entry.

At Canonical, we believe that Ubuntu is the ideal foundation for Hadoop clusters. Critically, it allows virtually unlimited scalability of Hadoop with no increase in licensing costs, and supports deployment on dedicated servers, on virtual machines, or in the cloud.

Because new versions of Ubuntu are released every six months, organisations can benefit from all the latest Hadoop features and toolsets faster. What's more, Canonical's close relationship with Hadoop developers at Hortonworks ensures that the next generations of the technology will be tightly packaged for Ubuntu.

Finally, Ubuntu provides out-of-the-box support for the commodity hardware that is typically used to build Hadoop clusters, helping minimise deployment costs and avoid frustrating compatibility issues. What's more, enterprise support is available for Hadoop environments running on Ubuntu from Canonical, the company behind the Ubuntu project.

If you do want more information on Ubuntu, how it works, what it offers, and how it can benefit your business today, please register your details at:

ubuntu.com/business/services/contact

Alternatively, you can download any of our white papers at:

canonical.com/whitepapers

1 http://hortonworks.com/