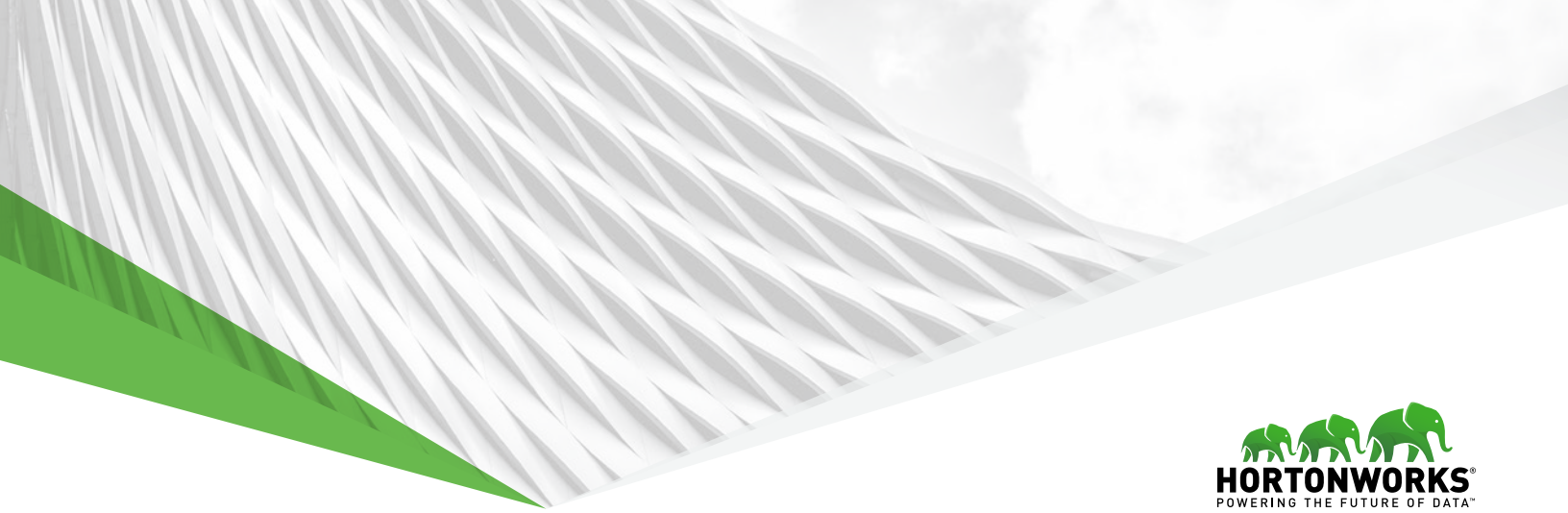

# WHY A CONNECTED DATA STRATEGY IS CRITICAL TO THE FUTURE OF YOUR DATA

A Future of Data white paper for CIOs, CTOs, IT & Data Strategists

A HORTONWORKS WHITE PAPER
MARCH 2016

# Contents


# Abstract

The advent of big data revolutionized analytics and data science and created the concept of new data platforms, allowing enterprises to store, access and analyze vast amounts of historical data. The world of big data was born. But existing data platforms need to evolve to deal with the tsunami of data-in-motion being generated by the **Internet of Anything** (IoAT).

Driven by consumer behavior, in order to personalize, sell, market and support, successful business need to go beyond understanding what has happened and focus on understanding and predicting what happens next. We need a fresh approach to get the best value out of our data. We need a better ability to get the tsunami of data-in-motion into the data platform while concurrently identify insights in real-time. We need a data strategy that is more enterprise-ready and future proof. This white paper examines the drivers and requirements for a fluid and interactive Connected Data Platforms strategy that handles both data-in-motion and data-at-rest.

# The Evolution of Data Platforms

Only a few years ago, organizations started creating data warehouses to improve their access and insight into historical information such as ERP systems, CRM systems, and system of record. This soon proved to be too expensive, too slow and not agile enough to handle new less structured and unstructured data such as log files, click-stream data, log files, and social media. Big Data and Apache™ Hadoop® were born to address these issues and the first data platforms started to emerge.

The ability of the data platform to store and analyze data at low cost and to handle all data types has allowed organizations to see new kinds of value chain emerge, such as speed and quality of web search, more effective web advertising, analysis of customer interactions and behaviors. This sparked the imagination of what was possible with data.

That value chain is real and continues to expand. According to McKinsey[1], if you're a Fortune 1000 company and you digitize an asset, you can grow its value by 5x. If you sell through digital channels you can increase productivity by up to 15%. If you digitize your business strategy you can realize a 3x increase in profit margins. This would be impossible without Apache Hadoop and the data lake. processing of large data sets to a fully-fledged data platform with the necessary services for the enterprise from Security to Operational Management and more.

1. http://www.mckinsey.com/business-functions/business-technology/our-insights/the-internet-of-things-the-value-of-digitizing-the-physical-world

# The Tsunami of The Internet of Anything Keeps on Coming

In the past, the world's data doubled every century. Now it doubles every two years. This flood of data is driven by the Internet of Anything, including more data from the Internet, mobile devices, server logs, Geolocation coordinates, Machine and sensors. With the advent of more connected 'things' such as wearables, sensors, artificial intelligence and so on it promises to continue to increase exponentially. This has put pressure on the data platforms to evolve to support this data-in-motion.

Today we are moving even beyond the wildest predictions of the growth of data being created, stored, retrieved and analyzed only a few years ago. With 5G the smartphone can potentially deliver a staggering 1Tbps[2] of data! In just two years from now there will be more devices than people on this planet. 35% of Americans Now Own at Least One Smart Device other than a Phone like thermostats, refrigerators, watches all delivering signals.[3] That's up to 6.4B of connected things, or 21B devices, by 2020.

The speed and flood of data means that the digital universe will grow from 4 zettabytes of data to 44 zettabytes this decade. 1.7 megabytes of new information will be created every second for every human being on the planet, 1/3 of it passing through the cloud.[4]

Data types are constantly changing too. Obviously, it's not just rows and columns any more; image data, all kinds of streaming data, geospatial coordinates, and time series. The one and a half billion monthly active Facebook users has represents more than 140 billion friend connections to be made, 265 billion photos uploaded, 62 million songs played 22B times.[5]

Gartner tells us that 32% of businesses who undertook a digital business transformation say their businesses are now digital businesses.[6] We believe at some point every digital business will also be a data business and data potentially their most valuable asset.

According to Forbes:

- 59% of businesses consider data and analytics to be "vital" to the running of their organizations, with a further 29% deeming it "very important".
- 69% say there is business case for investing in exploring new ways to add value through data projects.
- 83% say it is making existing services and products more profitable.
- 60% of businesses claim their data is generating revenue within their organizations.
- But 48% feel that their organizations have, in the past, failed to take advantage of opportunities to capitalize on their data.

The Internet Revolution opened up new frontiers to accelerate productivity, reduce inefficiency and waste, and enhance the human work and life experience. Whatever you call it, the current phase is accelerating this. New disparate technologies to solve each individual problem are becoming a dime a dozen, however data-at-rest and data-in-motion technologies that can interact with each other and share common operations, security and governance are key. These technologies are fueling the Connected Data Platforms.

2. http://www.trustedreviews.com/news/5g-researchers-crack-1tbps-data-transfer-at-uk-university
3. https://www.truste.com/about-truste/press-room/35-of-americans-now-own-at-least-one-smart-device-other-than-a-phone/
4. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
5. http://www.ge.com/docs/chapters/Industrial_Internet.pdf
6. http://www.gartner.com/technology/research/digital-business/

# Data Platforms Need To Evolve

Firstly, this tsunami of data is only increasing and data platforms need to evolve to meet the need. No longer can a data platform only deal with data-at-rest and batch operations. The platform needs to be connected to the Internet of Anything.

Connected Data Platforms need to handle distributed data across departments, servers and location. This data can be data-at-rest or data-in-motion. It needs to be done securely and cost effectively and take into account bandwidth and connectivity issues.

Connected Data Platforms need to go beyond Apache Hadoop and need enhanced data routing, transformation, and system mediation logic that is starting to emerge from projects like Apache NiFi[7] which supports real time ingestion, pattern detection, routing, and analysis.

Second, many organizations cannot realize the value of the data in their data platform because they need rocket science expertise to analyze and inspect the data. Distributed data processing engines like MapReduce or Apache Spark and other scripting and query capabilities such as Apache Pig and Apache Hive can be just too complex to use by the masses. New capabilities and tools like Apache Zeppelin are needed to make the Connected Data Platforms more accessible to mere mortals. In general we see the level of abstractions continue to rise to simplify and democratize analytics.

Third, data platforms need to evolve to provide capabilities to make it acceptable for enterprise use. Connected Data Platforms need to be enterprise ready, enterprise scale, provide predictable performance, support data encryption, security, data governance, HA, DR, operations and debugging. Apache Hadoop will be used in 100% of enterprises, and enterprises demand quality, uptime, and the peace-of-mind in knowing that support is available 24 hours a day, 365 days a year.

**APACHE ZEPPELIN**

Apache Zeppelin is a web-based notebook for agile analytic development. This open source tool provides a visual interactive experience for uncovering insights and sharing those insights with others.

7. https://nifi.apache.org

# Modern Data Applications Require Connected Data Platforms

Connected Data Platforms are needed to both capture perishable insights from data-in-motion while ensuring rich, historical insights from data-at-rest. The rules of the game have changed with the variety, volume and velocity of data, and the heightened expectations of data that has to be analyzed simultaneously. Understanding what happened is no longer good enough; we need to predict what happens next, and to do this we need platforms that can connect data-at-rest and data-in-motion. Autonomous self-driving cars are a great example; for them to hit the roads we need them not to hit the other cars and this requires a level of prediction of what is happening in real time, at speeds over 55mph based on historical data of positive outcomes that have happened in the past and continuous analysis of real time data coming from dozens if not not hundreds of sensors. The prediction is that by 2030 self-driving cars will likely be dominating the roads, but this cannot happen without a Connected Data Platform.

At a more practical level, the future of customer service, marketing and selling is via modern data applications. Now more consumers (especially millennials) want to access to your brand or services via smart apps versus talking to a human. If you stayed in a high end hotel lately, for example, its is likely they already starting to replace the concierge and the in-room phone with a tablet plus a data driven app. We call room service from the app! As consumers we expect (demand) a secure, personalized and in-context experience from that smart device app. To do this the businesses needs to analyze multiple types of data-at-rest and data-in-motion simultaneously in real time or near real time to understand context, personalize and to start to predict.

Today's leading-edge use cases are therefore based on modern data apps that can convert yesterday's impossible challenges into today's new products, cures, conveniences and life saving innovations and all of them require a connected strategy.

Another example is the use case for marketing automation which used to be based on analysis of cookie tracking or web behavior compared to our best guess at rational or emotional based personas, then scoring and automating our response via specialized online marketing tools. The modern use case is now the automatic recommendation engine that can match products to preferences in milliseconds based on web behavior or physical location or other context. What did you click on a few seconds ago? How does that map to the billion records of others who clicked on the same in the data platform? Or, when did you enter and when did you leave the store? What social media were you using at the time? Which product and beacon did you stand next to? What other offer did you respond to? What can we now offer you?

# A Real World Example: Progressive Insurance

Progressive Insurance is using a connected data strategy to reward safer drivers and improve traffic safety. You may have seen Flo advertising their Snapshot plug-in to capture driving detail for policyholders. Through a web app, customers can review their own driving and improve safety as well as Progressive.

With more than ten billion driver miles stored in Hortonworks, Progressive can effectively predict risk down to the individual driver and price their usage based policies, giving safer drivers deeper discounts.

Plus, as those driver miles stream in and accumulate in the Connected Data Platform, Progressive's predictive models get stronger and help with other parts of their business, like mining insurance claim notes for accuracy.

From this, Progressive has been able to achieve transformational business outcomes and dramatic cost savings from this. Snapshot and usage-based insurance drove $2.6 billion in incremental 2014 premiums!

# Six Requirements of Connected Data Platforms

We see these six requirements for a connected data strategy. Ask yourself, does your data platforms meet these requirements:

1 **Secure Data Ingestion**: Can you easily and securely ingest data from anywhere across the Internet of Anything and siphon off data quickly, detecting interesting patterns?

2 **Actionable Intelligence**: Can your platform provide real-time actionable intelligence based on both data-in-motion and data-at-rest?

3 **Distributed Connectivity**: Can you connect, relate, or put in context all forms of data to provide 360 degrees of context that allows you to predict and personalize data for your modern data applications across all your channels?

4 **Mere Mortal Friendly**: Does your platform provide data science tooling that make this easier for data scientists and professionals to find interesting patterns and gain actionable intelligence?

5 **Future Proof**: Is your platform built on 100% open source technology allowing you to capitalize innovation everywhere economy?

6 **Enterprise Ready**: Do you have enterprise-ready capabilities for security, high availability, disaster recovery and so on? Is the data secure? Is it access controlled? Are all compliance regulations taken care of? Is activity tracked with an audit trail? Is data controlled and managed through a lifecycle? Is your platform supported by experts?

## What Hortonworks is Doing

Our technology strives to adheres to all these requirements. We view Connected Data Platforms as the most effective open approach to managing and gaining insight from data-in-motion and at-rest.

The Hortonworks Data Platform or HDP™ addresses the complete needs of data-at-rest. HDP is architected, developed, and built completely in the open, providing an enterprise ready data platform that enables organizations to build Modern Data Applications. With YARN as its architectural center it provides a data platform for multi-workload data processing across an array of processing methods — from batch through interactive to real-time, supported by key capabilities required of an enterprise ready platform — spanning Governance, Security and Operations.

Part of our Connected Data Platforms strategy is Hortonworks DataFlow (HDF™). HDF addresses the needs of "data-in-motion" such as real-time data streaming capabilities and is a cornerstone technology for the Internet-of-AnyThing (IoAT) and data provenance use cases.

Combined together HDP and HDF provide the industry's most open, innovative and enterprise ready Connected Data Platforms.

## About Hortonworks

Hortonworks is a leading innovator at creating, distributing and supporting enterprise-ready open data platforms. Our mission is to manage the world's data. We have a single-minded focus on driving innovation in open source communities such as Apache Hadoop, NiFi, and Spark. Our open Connected Data Platforms power Modern Data Applications that deliver actionable intelligence from all data: data-in-motion and data-at-rest. Along with our 1600+ partners, we provide the expertise, training and services that allows our customers to unlock the transformational value of data across any line of business. We are Powering the Future of Data™.

**Contact**

For further information visit
For more information,
visit www.hortonworks.com

+1 408 675-0983
+1 855 8-HORTON
INTL: +44 (0) 20 3826 1405

HORTONWORKS
POWERING THE FUTURE OF DATA™