CASE STUDY

Thomson Reuters Cuts Time for App Development on Hadoop with CDAP

the answer company™
THOMSON REUTERS®

Client

Thomson Reuters is the world's leading source of news, data and information that help professionals find trusted answers and power global markets. Thomson Reuters operates in more than 100 countries and reaches over one billion people per day across the world.

Challenge

With the goal of enhancing their customers' ability to find trusted answers to key questions in their industries, Thomson Reuters decided to make the company's most valuable asset, data, more accessible and discoverable in a large-scale data lake platform on Cloudera CDH. Co-location of content would drive innovative product ideas with the ability to find and use data from disparate sources across the business. However, onboarding new Hadoop developers took up to four months before they were productive at building and deploying custom applications at scale, a challenge which Thomson Reuters recognized and wanted to accelerate.

Solution

Following a proof of concept designed to benchmark and demonstrate the productivity gains that Thomson Reuters could expect from using the Cask Data Application Platform (CDAP) in their environment, the engineering group realized that CDAP could significantly shorten some teams' learning curve when working with Hadoop and Spark. Further, CDAP would help them scale their data lake project much faster.

Benefits

- By significantly reducing the time spent on operational tasks such as creating centralized logging and auto-scale tools, CDAP reduced Thomson Reuters' time to build applications, accelerating time-to-market and functionality of this large-scale data lake
- Using CDAP, it took new members of the Thomson Reuters platform engineering team only about a month to be productive using Hadoop, compared to the traditional 3-4 months of ramp.
- Due to CDAP's standardization and integration layer, which enables the creation of reusable and portable components, the platform engineering teams are now in a great position to support the growing big data needs across the various business units that these teams offer core platform and infrastructure services to.

Enabling Customers to Find Trusted Answers to their Questions

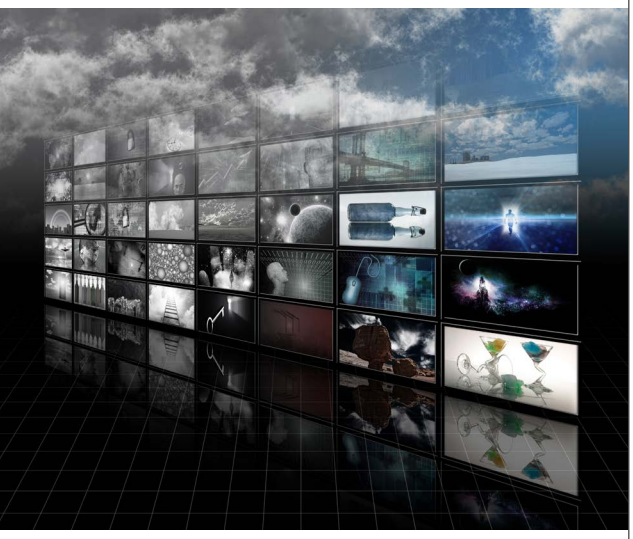
Thomson Reuters, "The Answer Company", is a global enterprise that provides instant news and intelligent information to the world's business and professional markets. It provides professionals in many industries – financial and risk, legal, tax and accounting, and media – with the intelligence, technology and human expertise they need to find trusted answers and make critical decisions.

With the goal of continually enhancing how customers find trusted answers, Thomson Reuters wanted to make the company's most valuable asset – data – more accessible and more discoverable to customers to drive innovation and create new revenue opportunities. The group decided to build an enterprise content ingestion and content management platform that would allow their different lines of business to design customized data packages for their customers, drawing on the wealth of information from across the company – a modern, large-scale data lake on Cloudera CDH. However, building custom applications that would reliably ingest, process and serve petabytes of structured and unstructured data was no small feat. Ingesting data with Sqoop, writing MapReduce and Spark jobs, scheduling Oozie workflow jobs, and making it all work at scale was a big challenge for a team that had only a limited number of experienced Hadoop developers.

> "For the proof of concept, we decided to do it two ways: The hard way, developing directly on Hadoop, and also another way of using the CDAP framework. Soon, we could see gains of efficiency in onboarding developers much faster on the CDAP framework."
>
> *– Vsu Subramanian, Vice President of Platform Engineering*

Thomson Reuters recognized what the Cask Data Application Platform (CDAP) could do to help accelerate functionality and eventual time-to-market for the data lake. CDAP, an open source, enterprise-grade unified integration platform for big data, is a layer of software that runs on top of all leading Hadoop distributions both on-premises and in the cloud, and allows organizations to accelerate the development, deployment and operations of data-driven applications. By significantly reducing the amount of operational coding and tool integration, CDAP cuts the average time to implement straight forward ingest solutions by more than 60%.

CASK



CDAP provides time-saving abstractions and pre-integrations, a complete testing and packaging environment and comprehensive data integration capabilities. These broad capabilities provide significant, rapid value for enterprises in a complex, data-heavy environment, such as Thomson Reuters, slashing the time to take big data projects from prototype to production. “Thomson Reuters’ platform and core infrastructure engineering teams provide reusable core platform services that can be leveraged across all our businesses”, said Sandy Martin, Senior Director of Enterprise Content Platform Engineering at Thomson Reuters. “The shared operational functionality that CDAP provides supports the mission of our central technology organizations.”

Questions Asked, and Questions Answered

Working closely with Cask and Cloudera, Thomson Reuters developed a proof of concept to see if they could accelerate time-to-market for the new content platform, its suitability as a platform for developing custom big data applications, and the benefits derived from the reusability and portability of components built with CDAP. The proof of concept was also designed to benchmark and demonstrate the productivity gains that Thomson Reuters expected from using Cask Data Application Platform (CDAP) in their environment.

During the proof of concept, Thomson Reuters did a side-by-side comparison. The engineering group’s ability to develop applications with CDAP was measured against developing the same applications directly on top of Hadoop. To that end, two teams were created – one composed of half a dozen of Hadoop experts with some prior experience working on the project, which was asked to continue developing directly on Hadoop. And a second team, which was half the size, and composed of developers that were skilled, but with no prior Hadoop or project experience; the second team was asked to develop the applications using CDAP. “For the proof of concept, we decided to do it two ways”, said Vsu Subramanian, Vice President of Platform Engineering at Thomson Reuters. “The hard way, developing directly on Hadoop, and also another way of using the CDAP framework. Soon, we could see gains of efficiency in onboarding developers much faster on the CDAP framework.”

> “Within two weeks, the CDAP team, who were new to Hadoop, had caught up with our core Hadoop developers; they were up and running, developing code, deploying, and we started to see results quickly.”
>
> *– Vsu Subramanian, Vice President of Platform Engineering*

After only three days of training, the Thomson Reuters team with no Hadoop expertise could develop applications with CDAP about five times faster than the team not using CDAP – a clear advantage that demonstrated faster time to value via increased development productivity. “Within two weeks, the CDAP team, who were new to Hadoop, had caught up with our core Hadoop developers”, said Vsu Subramanian. “They were up and running, developing code, deploying, and we started to see results quickly.”

"Without CDAP, each team would have had to build operational tooling," said Sandy Martin. In her opinion, the proof of concept demonstrated the value that accelerating learning time and scalable tooling with CDAP could bring to the projects like the data lake. "The alternative to using CDAP for these apps was to implement a workflow for each need. Then we would have to find a way to scale that workflow dynamically. Developing scaling solutions for every workflow isn't sustainable at scale", she said.

Sandy continued, "Other solutions in the market typically focus on only tackling tactical data integration problems, but don't help with the actual applications or their deployment into production. CDAP lowers the barrier to entry to working in a Hadoop environment. CASK management of the code base as open source gave us the opportunity to set direction for the features built over the year."

Scaling up to Offer More Answers – Instantly

As a result of the successful proof of concept, Thomson Reuters decided to adopt CDAP to build out its content platform, bringing content from many sources – RDBMS, but also data feeds from different data distribution frameworks – together in this single, Hadoop-based platform. Moreover, the platform engineering group determined that CDAP should be the integration framework of choice for any new development of Hadoop and Spark applications at Thomson Reuters. This was based on the key findings from the proof of concept, which included increased development productivity with CDAP, compelling visualization of workflows delivered by CDAP, and the extensive and powerful API's native to CDAP.

"The alternative to using CDAP for these apps was to implement a workflow for each need. Then we would have to find a way to scale that workflow dynamically. Developing scaling solutions for every workflow isn't sustainable at scale."

– Sandy Martin, Senior Director of Enterprise Content Platform Engineering

The above capabilities help Thomson Reuters to operate a large-scale data application development and operations environment where workflows are created and leveraged across all of Thomson Reuters' business units.

"Normally, when dealing with Hadoop, you have to learn how to interact with HDFS, HBase, YARN, Sqoop, Oozie, etc. and how to stitch these pieces together; that can easily take many months. In contrast, CDAP provides abstractions for these low-level API's, which have already significantly accelerated time-to-value from these technologies for us now, and will save us even more time as our organization rolls out future big data projects to the business," said Sandy Martin.

> "CDAP lowers the barrier to entry to working in a Hadoop environment. CASK management of the code base as open source gave us the opportunity to set direction for the features built over the year."
>
> – *Sandy Martin, Senior Director of Enterprise Content Platform Engineering*

The content platform built with CDAP on top of Cloudera has been running successfully in pre-production at scale, demonstrating its ability to meet rigorous stability, reliability and operational requirements. CDAP will be a critical factor when operationalizing the content platform for customer trials later this year. "It's one thing to develop code, it's a whole other thing to run it in an enterprise way – which means operations and DevOps teams know how to deploy and troubleshoot code, meet SLA's, meet data quality requirements," said Sandy Martin. "We were able to build company-standard logging and monitoring mechanisms around CDAP applications allowing the CDAP applications to run in an enterprise fashion."

About Thomson Reuters

Thomson Reuters is the world's leading source of news and information for professional markets. Our customers rely on us to deliver the intelligence, technology and expertise they need to find trusted answers. The business has operated in more than 100 countries for more than 100 years. Thomson Reuters shares are listed on the Toronto and New York Stock Exchanges. For more information, visit www.thomsonreuters.com.

About CDAP

The first unified integration platform for big data, Cask Data Application Platform (CDAP) lets developers, architects and data scientists focus on applications and insights rather than infrastructure and integration. CDAP is open source and accelerates time to value from Hadoop through standardized APIs, configurable templates and visual interfaces. With a radically simplified developer experience and a code-free self-service environment, CDAP enables enterprise IT to broaden the big data user base and seamlessly integrates with existing MDM, BI and security and governance solutions.

About Cask

Cask makes building and running big data solutions on-premises or in the cloud easy with Cask Data Application Platform (CDAP), the first unified integration platform for big data. CDAP reduces the time to production for data lakes and data applications by 80%, empowering the business to make better decisions faster. For more information, visit the Cask website at cask.co and follow @caskdata.

Cask Data, Inc. | 150 Grant Avenue, Palo Alto, CA 94306 | 650.469.328 | cask.co