

HOW RAILING IS HELPING TO BUILD A SMARTER RAIL NETWORK LEVERAGING AUTOMATION AND BIG DATA



lished in 1999 to provide IT services for the Association of American Railroads (AAR), Railinc is the railroad industry's most innovative and reliable resource for IT and information services. Today, as a wholly owned subsidiary of the AAR, Railinc supports business processes and provides business intelligence that helps the freight rail industry increase productivity, enhance operational efficiency, and increase their return on investment in assets.

In recent years, rail industry participants have recognized the opportunities big data presents. In response to the critical need for actionable data, we are working with Class I, short line, and regional railroads as well as other Railinc customers to capture huge volumes of data across diverse points in the rail network and help our customers:

- Track shipments across the North American freight rail network
- Achieve efficiencies around railcar repairs, car hire and other rail operations
- Monitor the health of equipment to ensure the safe movement of freight
- Better manage traffic to keep railcars moving

Railinc is now the largest, single source of real-time, accurate interline rail data for the North American railroad system. That data is empowering our customers to drive efficiency, manage costs, and improve the health of the North American rail network.

To continue meeting our customers' data needs, we've embraced big data and replaced a previous proprietary data warehouse with an open-source environment that offers greater flexibility at a lower cost. Two years ago, we began the move to Hortonworks Hadoop as the framework for storing, processing, and managing the massive volumes of data we handle today and support even greater data volumes in the future.

Control-M is a vital part of our big-data strategy. Railinc has used Control-M for 11 years to schedule and monitor complex batch processes across multiple platforms and applications. Control-M for Hadoop allows us to develop, schedule, and monitor Hadoop batch processes using the same familiar interface we use for our other workloads.

Big Data Brings Unprecedented Visibility

The North American rail network is growing increasingly smarter as railroads implement advanced technology such as intelligent sensors positioned alongside tracks. These sensors provide data such as location and movement information that helps customers manage their fleets, track their equipment, view ETAs, efficiently coordinate the movement of millions of railcars and time the delivery of cargo down to the hour. Still, other detectors monitor the physical condition of rolling stock, enabling railroads and car owners to detect issues such as a bad brake or a wheel with a flat spot. These data enable Railinc to provide advanced warning to schedule repairs before a minor issue becomes a costly repair.

The volume is staggering. We're capturing data from more than 40,000 locomotives and 1.6 million railcars traveling across 140,000 miles of track. The data come from equipment belonging to 1,700 different rail car owners, 560+ local and regional railroads, and seven Class I railroads. Our data warehouse already contains 50 terabytes of data from disparate sources and we expect that volume to increase nearly 100% over the next few years.

Railinc industry applications leverage these data to enable customers to operate more efficiently and economically. Our car hire applications support activities around the fees charged and paid for

the usage of rail equipment, enabling higher equipment utilization and improving payment accuracy. Traffic management applications such as our Clear Path™ System facilitate the movement of trains through the Chicago Terminal, the busiest rail gateway in North America.

Control-M Helps Keep Big Data Flowing

To support our industry applications, we must gather huge volumes of data every day from many sources, move it through various systems for analysis and translation into actionable information, and generate and distribute reports to our customers. The workflows that get the data where it needs to be when it needs to be there are highly complex with numerous dependencies.

That's where Control-M comes in. It simplifies the creation of even the most complex workloads. Using the graphical interface, I can literally draw the dependencies among jobs, so I can ensure that prerequisite processes in a sequence are completed before the next process in the sequence is started.

Perhaps the most significant benefit is that the solution isn't tied to a single technology. When the big-data team added Hadoop, Control-M was a natural fit, giving us the same kind of visibility into and control of Hadoop jobs that we experience on other platforms. The scheduling staff didn't have to learn a special scheduling tool for Hadoop. We use the same interface to schedule workloads on all of our platforms and we have full visibility into the hundreds of jobs that run every night.

Control-M Batch Impact Manager lets us monitor Hadoop jobs without having people sitting in front of a console 24/7. If the solution detects potential delays or failures, it alerts us immediately. Our customers rely on us to meet those SLAs. Delayed reports on the Chicago Terminal, for example, could affect rail operations not only in the Chicago area, but throughout North America. Batch Impact Manager gives us an intelligent, proactive approach to keeping processes—and trains—running.

Another plus for Control-M is that it supports multisite environments. The critical nature of our big-data environment makes it important to have a second site for disaster recovery purposes. So we created a primary and secondary site for our Hadoop environment.

However, we aren't limiting the second site to DR activities. We need the flexibility to do any job on either site. Control-M can easily talk to both sites, enabling us to ETL to either site, manipulate data, create views, and move processes back and forth between sites as necessary and ultimately replicate IT services and data on both sites. In the future this multisite capability will help with load balancing, enabling us to keep up with the increasing demand for big-data reports.

Conclusion

The transition to Hadoop has added significantly to the number of jobs we run and Hadoop workflows now account for about one third of all our batch processes. The orchestration that Control-M performs to get data where it needs to be at the time it needs to be there is critical to the success of our big-data efforts.

We simply couldn't do the job without Control-M.

For more information on Control-M for Hadoop, [click here](#).