CONTINUOUS DELIVERY MODEL FOR BIG DATA PROJECTS

The continuous delivery model has been thoroughly embraced by the technology industry. Companies like Amazon, HP, and Spotify have migrated their development processes to take advantage of this model for delivering applications. This white paper explores the benefits, challenges, and techniques for applying the continuous delivery model to big data projects.

What is Continuous Delivery?

Continuous Delivery is an operational approach that allows teams to get changes of all types into production, or into the hands of users, safely and quickly in a sustainable way. The goal is to make deployments of the system a routine operation that can safely be performed on demand.

The key to this is insuring that the code is always in a deployable state and completely eliminates the traditional testing and deployment phases of conventional software workflow such as “dev complete” and code freezes.

It is based on five simple principles.

- **Build quality in**
- **Work in small batches**
- **Computers perform repetitive tasks, people solve problems**
- **Relentlessly pursue continuous improvement**
- **Everyone is responsible**

Benefits to Big Data Projects

While the Continuous Delivery model has received tremendous uptake in traditional application development processes, data analytic projects have remained stubbornly stuck with the waterfall project management approach.

The biggest change afforded by continuous delivery is that teams are able to get working software in the hands of users quickly and iterate often.

Most analytic projects evolve as users become familiar with the data. If the business can start seeing aspects of the analytic early, they will come up with entirely new ideas on directions that the data can take them. The continuous delivery model embraces this working relationship rather than waiting until the end of a long project for this feedback and resisting scope changes throughout.

This model also draws business users deep into the process. By giving them working software early and often they gain ownership over the data and the process and become partners in the endeavor.

Finally, relentless focus on automated testing helps to build quality into the process. Unit tests and frequent deployments to users help catch bugs early, before they impact more of the system. User feedback also helps build confidence in the analytics so they can be put to use in the business.

Challenges to Implementing Continuous Delivery

Naturally, implementing a change as profound as continuous delivery is not without challenges. Here are some typical examples, along with techniques that STA Group use to overcome them.

Scarcity of Automated Testing Tools in the Big Data Space

Unlike the Java or Node application space, there are limited tools for performing automated tests on big data applications. STA has years of experience in Test-Driven-Development and have created toolkits to help provide the level of code coverage we feel necessary for these projects. We are working with an important Open Source project to build unit testing tools for Spark which we believe will help the entire big data community.

Fragmented Teams

Analytic projects are complicated in how they often span so many disciplines and business areas. This leads to fragmentation and delays as tasks are handed off between development teams. STA Consultants are experienced Scrum practitioners and help our clients build multidisciplinary teams that cut through these artificial barriers and keep the project focused on delivering business value.

Complex Dependencies in the Data

Most analytic projects involve layer upon layer of data extraction, transformation, modeling, and further transformation. Finding quick wins and paths that deliver immediate business value can be challenging. It takes skills in understanding the data architecture and experience in crafting user stories to create a backlog that will deliver on the benefits of continuous delivery.

Lack of Suitable Continuous Integration Environments

One of the keys to implementing this model is the ability to perform automated tests of the evolving software and quickly deploying the system to production. The whole big data ecosystem is very complicated and cumbersome to utilize in a continuous integration pipeline. We have invested heavily in engineering containerized versions of the big data environment, as well as elastic cloud-based deployments. We are able to create cost effective, integrated build, test, and production environments that meet the demands of Continuous Delivery. Furthermore, we are leading experts in the growing field of Analytic Ops, and have pioneered tools for managing the deployment of new models to production environments.

Contact Us

If you are interested in learning how the continuous delivery model can improve your big data project effectiveness, please contact Mark Harrison, Managing Director, at 630-263-6371 or reach out to him at mark.harrison@stagrp.com.