The importance of DataOps: Why it matters and how to get started

Data management has been facing challenges since the early days of web development and IT. The nature of the challenges has been changing from procurement to storage, to high volume storage, and from transaction to deriving insights, to making the whole process fast and efficient.

Gaining agility to imagine real-time insights from fast burgeoning data in an automated manner is the need of the hour and the focus of many experts in the field of data today. This focus has resulted in the emanation of DataOps.

DataOps draws parallel with a similar yet different practice in software delivery and SDLC management practice called DevOps, which uses large scale automation to quicken the software release cycles and improve the quality of output. DataOps is the amalgamation of various processes and practices supported by relevant technology that uses automation to boost agility with respect to data and insights. It thus augments the speed and quality of data, information and insights, and supports continuous improvement culture of businesses.

Data challenges and the DataOps code

DataOps as a practice tries to address some of the data management challenges to ensure quick and quality analytics for the end users.

Data and related insights tend to lose the value that they possess with each tick of the clock. The requirements in the data and analytics space fluctuate: sudden changes in requirements are common. Also, the constant quest to ask more questions and counter questions from data keeps the analytics process on its toes.

The number of data pipelines have grown with requirements from data analysts, scientists, and data-hungry applications resulting in data silos with no connect to other pipelines, data sets and data producers. The data resides in different systems and platforms; gaining access and control over different systems and identifying the right data becomes a daunting challenge.

Bad data quality causes lost credibility in the entire analytics setup and leaves the whole program in jeopardy. The data formats may be different in different systems depending on the datatypes and schemas. Data errors can creep in due to numerous events like duplicate entries, schema change and feed failures. These errors can be difficult to trace and deal with. Also, there are constant updates in the data pipeline in terms of say, schema changes, added fields new/updated data source etc. Making and validating these changes is a constant and time-consuming task.

Manual processes in the data pipeline from integration to testing and analytics are overbearing and time consuming. Some teams work long hours to make sense out of the data analytics efforts, some commit less and take more time to be watchful in their approach, while others try to be quick by compromising on parameters. So, overcoming the data-related challenges requires not just tools but also changes in the underlying processes that handle analytics.

DataOps helps overcome the hurdles and complexities and deliver analytics with speed and agility, without compromising on data quality. It derives inspiration from the practices of Lean Manufacturing, Agile and DevOps. It emphasizes on integration, cooperation, collaboration, communication, measurement and automation between data engineers/ETL engineers, IT, data analysts, data scientists, IT and quality assurance. Thus, it focuses on getting fast insights by leveraging the interdependence of every chain of the analytics process by focusing on people, process and technology.

Implementing DataOps

DataOps can be accomplished by accommodating a few changes to the as-is process being used in the data setup. Automated testing can help save a lot of time and curb manual efforts. The data should be constantly monitored for quality issues and test cases should be added incrementally for every added feature or change to the input data records. The corresponding errors in the statistical process control should be projected as alerts to the appropriate stakeholders to keep the quality of data high.

Use of Version Control Systems wherever possible will help keep artefacts in a structured manner, enhance reusability and support multi-developer environments. Essentially, each step of the data analytics pipeline is accomplished by tools which are nothing but code i.e they produce underlying scripts, config files, parameter files etc. which if versioned in version control systems like GIT, SVN, customized version control tools/systems etc. will enhance reusability and reproduce a state in the pipeline when required. Also, multiple developers like ETL developers or analysts creating models can work on the same piece of code in parallel through branching the trunk and merging it back

by testing and incorporating the changes. This will save development time and speed up analytics.

Isolated environments will prevent chaos in the production database systems and Cloud environment can fulfill the separate data needs of developers. A data engineer making changes to data can cause issues if done in production directly or an analyst creating new model can be puzzled with constantly-changing dataset. Thus, isolated environments can be provisioned on the go. Code, configurations, parameters, tools or environments in the data analytics pipeline can be reused by effectively utilizing containerization technologies like Docker containers. Also, certain complex operations/steps in the pipeline can be enabled through containers (calling custom tools/API, downloading files, running a script e.g. python/shell script execution, FTP file to another system). This will enable the teams to execute the container wherever required instead of performing them from scratch and deploy it easily without understanding the underlying tools and configurations.

Flexibility is an important characteristic that must be present in the data analytics pipeline. The changes in approach or control of the process must be with the data engineers/analysts to make minimal changes/tweaks and get desired results. E.g. which version of datasets to use, which environment to consider, what condition to use for processing and so on. This can be effectively realized through use of parameters, by which the changes can be adjusted and incorporated on a case-to-case basis with minimal manual changes.

DataOps in action

A DataOps streamlined process encompasses tool chain and workflow automation when data enters the systems from sources and keeps on changing with time to feed the downstream systems for transformation, models, visualizations and reports. This can be taken as the production environment that directly exploits existing workflows, tests and logics to derive value and keep the quality of data in check at all times. So, the code or toolset remains constant and data keeps on changing and updating the downstream, keeping the insights live and active. This can be done with a good orchestration tool with automation as its credo.

Another activity that happens simultaneously is the generation of new code, tests, models and features to the existing code/tools that play with the data. This can be well handled by using fixed

datasets and containerized environments with parameters and versioning to enable developers, testers and other stakeholders speed up the changes to production. Thus, it speeds up analytics and strengthens the feedback mechanism of the pipeline (See Figure 1).

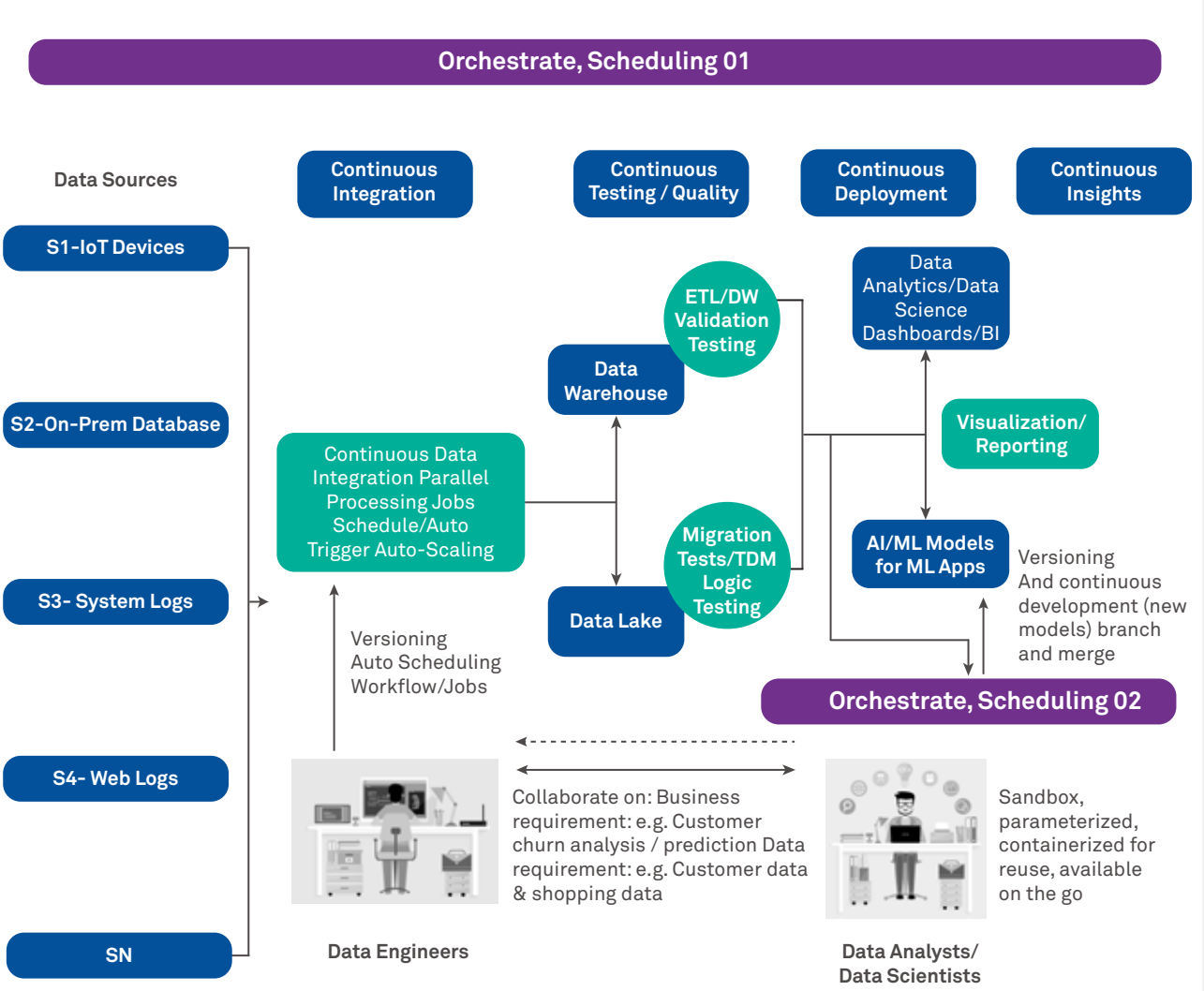

Figure 1: A unified approach to analytics with DataOps

Conclusion

Imbibing promising changes is a cultural phenomenon. A culture of reduction in timelines and improvement in quality that scaffolds constant improvement in metrics is a must-have for bringing about tangible gains out of a practice like DataOps. Every task/step in the process needs to be evaluated in terms of how automation and intelligence could sink in to make it better. The culture of constant improvement in both data and insights in terms of quality, agility and collaboration could be a right step in the direction of DataOps.

Continuity in efforts is required as there is no one thing that could make an organization DataOps-compliant. An understanding and sense of collaboration between the stakeholders (data engineers, analysts, scientists or stewards) involved in every step is required to ensure a streamlined process. Constant promotion, transformation, release of data should be coupled with constant application of models to derive insights and simultaneously improve analytics on a regular basis. Also, the analysts and developers should get a constant feedback on the quality and profile of data and insights which would quicken the process of taking corrective measures to ensure a robust process. This will be the direct implication of getting in DataOps in the scheme of things.

About the author

Pratyaksh Arora, Consultant, Information Management – Data, Analytics & AI, Wipro Limited.

Pratyaksh has experience in the field of DevOps, Information Management and Analytics. He holds an MBA from the Department of Management Studies, Indian Institute of Technology, Delhi and finds keen interest in new age practices like DevOps, DataOps and Data Management in general.

Mohan Mahankali, Principal Architect, Information Management - Data, Analytics & AI, Wipro Limited.

Mohan has 20 years of business and IT experience in the areas of Information Management and Analytics. He is the co-owner of a patent in Data Management & Governance awarded by USPTO (United States Patent and Trademark Office).

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035, India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 175,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**

IND/BRD/AUG 2019-JUL 2020