International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.23, August 2019– www.ijais.org

DataOps in Manufacturing and Utilities Industries

Prabin Ranjan Sahoo
Enterprise Architect
Manufacturing & Utilities Business Group
Tata Consultancy Services

Anshu Premchand
Agile & DevOps Practice Lead
Manufacturing & Utilities Business Group
Tata Consultancy Services

ABSTRACT
The concept of DataOps and its adoption across industries is gaining momentum. This paper draws a parallel between DataOps and DevOps concepts. It then focuses on the relevance of DataOps in manufacturing and utilities industries. The paper then outlines the dataOps process and platform as well as the data challenges in manufacturing & utilities industries. Various DataOps strategies for these industries are also discussed along with the importance of adoption of advanced analytics via DataOps for achieving business benefits.

General Terms
DataOps, models, monitoring, operations, archive, storage

Keywords
DataOps, utilities, data pipeline, DevOps, manufacturing

1. INTRODUCTION
Data is growing at an exponential speed because of the internet. It is envisaged to grow even faster with advent of autonomous vehicles, connected devices and the like in the digital market. IoT (Internet of Things), IIoT (Industrial Internet of Things) and edge devices are generating huge amount of data on a daily basis. While data generation is spontaneous, its consumption is not. With data piling up every day, fresh ways of thinking are required to accelerate its consumption and usage. From a business perspective, monetization through building powerful analytics is highly desirable.

A large number of organizations in manufacturing and utility industries are already working towards ways to monetize this data and / or use it to create insights beneficial for business. Manufacturing industries are continuously trying to improve operational efficiencies with the help of analytics. In addition, the autonomous vehicles need advanced analytics for their day to day operations – decisions such as choosing the best route, sensing unfamiliar situations to avoid accidents and predicting traffic and so on.

The utility industries are also not far behind. Smart metering and smart grid kind of technologies require & produce a lot of data. In power industry for example, such solutions require predictive analytics for better power distribution to make it profitable both for consumer and the industry. A lot of research is currently underway in the areas of advanced data analytics, big data analytics, and data science and the like to use this data for not only furthering business but also improving margins.

The research focus although is more on development of predictive analytics to create quick business values. The outcome is mostly point solutions, repetitive tasks, lack of synergy across product lines, groups, and inefficiencies [1]. This paper highlights how dataOps can bring revolutionary changes to business in the analytics space by use of dataOps. DataOps can eliminate the inefficiencies, creates opportunities for collaboration, and promote reusability to reduce operational costs, quick time to market by adopting scientific and disciplined approaches.

2. DATAOPS TO THE RESCUE
DevOps is a proven methodology to aid agile software development. This methodology holds good for any software services and product development [2]. DataOps is a derived concept out of devOps. Industry experts define dataOps as application of devOps to data - how effective data operations can be when devOps concepts are applied to data for managing, and deriving analytics. Figure 1 shows dataOps an expansion of devOps. There are few new stages marked with (✔) in Figure 1 which are required for achieving the overall effectiveness from data analytics point of view. These points are discussed in the following sections.

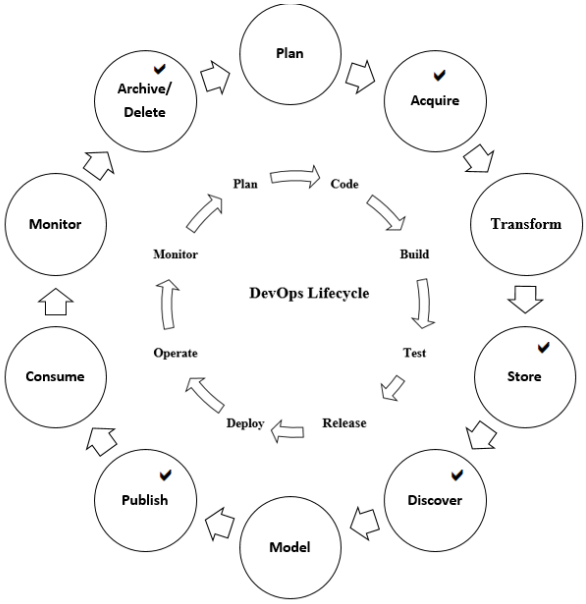

Figure 1: DevOps & DataOps

DataOps is devOps based and process-oriented methodology used by data and analytics teams to improve the quality of data analytics and reduce cycle time of the same, generally, to aid better business decision making, improving profits, furthering business and monetizing data available with an organization.

DataOps aids removal of data silos, improves backlog management & helps with improved data quality to ensure automated and faster data pipeline which in turn delivers business value continuously. DataOps enables continuous

deployment of data science models and makes the entire supply chain of data more repeatable, productive, agile and self-servicing while allowing the process to be rigorous, reusable, and automated for creating pipelines and applications for data.

The key components of a DataOps Platform can be summarized as follows:

2.1 Data Pipeline Orchestration

Data pipeline orchestration include the steps in the data analytics production such as data access, integration, model and visualization will need to be a directed graph-based workflow.

2.2 Assurance Automation, Production Quality & Monitoring or Alerts

Assurance automation focusses on automatic testing and monitoring of production quality of data and artefacts in the data analytics production process. The code changes will also need to be automatically tested during the process of deployment. Continuous monitoring and alert management solutions are also included here.

2.3 Continuous Deployment and Creating a Sandbox for Development

Continuous deployment focusses on movement of code as well as configuration continuously into production from development environments in a seamless manner. Also, there needs to be a rigor in research which necessitates creation of sandboxes such that data engineers can experiment and conduct proof(s) of concept as needed.

2.4 Deploying the Data Science Model

Data science teams may create development environments which can be reproduced. These environments are near-mirrors of the production environments to ensure bug free movement of the data science models into production.

The topmost priority for DataOps is to ensure customer satisfaction by ensuring that insights valuable to business are delivered quickly and continuously [3]. DataOps also focusses on accuracy of data and robustness of systems, applications and models in use.

Continuously managing change to generate value for business is a key focus area for DataOps [4]. DataOps thrives on regular and frequent interaction amongst customers, operations and data analytics teams.

Self-organizing teams is one of the finest principles of mature agile organizations, and it applies to DataOps as well. DataOps focusses on creating scalable, repeatable and sustainable processes. The Data teams also focus on 'continuous improvement' to continuously generate value for business.

One of the prime drivers of success would be end to end orchestration for the entire process [5]. It is also imperative to allow the team to work on ready & easy to use sandboxes.

DataOps has a continuous focus on monitoring and improving quality as well as performance.

DataOps also focusses on reuse and cycle time improvement, similar to devOps.

3. THE DATAOPS PROCESS

The dataOps process has six significant steps that require attention for eliciting right benefits from adoption of dataOps.

These steps, as outlined below are, business requirement planning, data acquisition, data transformation, data repository management, data modelling and insight publication.

3.1 Business Requirements Planning

The plan phase of devOps can be mapped to the plan phase for business requirements in the context of business analytics which includes the identification of consumers of data, producers or sources of data, data acquisition process, data model, and the overall process.

For manufacturing and utilities industries, IoT (Internet of Things) devices are primary sources of data. The sensors attached to machines and edge devices send large volumes of sensory data to a central hub or are locally stored and then uploaded into a central system.

3.2 Identification of Business Consumers

The consumers of business decide the requirements for the data analytics process in dataOps. As the business dynamics change with time so do the requirements for data analytics as evolving situations may demand new business insights.

This process needs to be done iteratively, that is why devOps principle best fits into this phase. It is also imperative to note that dataOps is not *do once and forget for a lifetime* task. The principle of dataOps is to allow the business and information technology teams to work together to draw trustworthy insights from business data from disparate sources quickly and efficiently and allow this process to improve iteratively.

3.3 Defining Scope and Objectives

The scope and objectives with regards to using vast abundance of data available with a firm to create insights that can be leveraged by the business can be clearly defined for an initial system. But large organizations, tend to be amalgamations of several business groups and it would be difficult to take everything into consideration. The goal will be to leverage all available and ever increasing data in a trustworthy, efficient and cost effective manner using dataOps principles. Figure 2 shows dataOps methodology to build analytics.

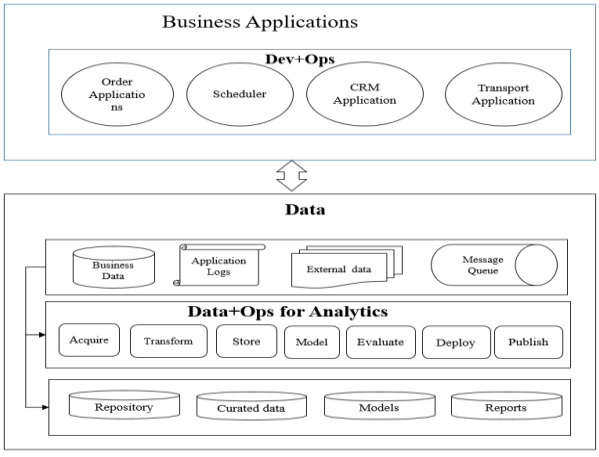

Figure 2: DataOps Methodology to build Analytics

3.4 Data Acquisitions

3.4.1 Data Source and Target Identification

In the data acquisition stage, external & internal data sources and targets need to be identified along with the challenges specific to data acquisition and disbursement processes. Some challenge areas can be related to security, accessibility, medium of data & data download time estimation and so on.

For example, if the data is either batched or resides in cloud or if the target is cloud then factors such as the available bandwidth and encryption mechanism need to be considered.

3.4.2 Data Source onboard

Each data source identified needs to go through an on-boarding mechanism. The on-boarding mechanism should meet certain criteria related to standardized sanity checks such as security, encryption, and data type & so on. These can be validated by using automated scripts. Data also needs to be consolidated in a data source that is under the control of the data analytics teams before it can be utilized. Such a source is sometimes referred to as the golden source of data.

3.4.3 Identification of tools

Based on the sources and type of data acquisition requirements, tools can be identified to create a dataOps pipeline. For example, real-time streaming would require a different set of tools as compared with a batch download. For enterprise use, both approaches can be considered. While selecting the tools, the capabilities of each tool should be evaluated to fit purpose and context.

3.5 Data Transformation

3.5.1 Current Challenges

Current data transformation techniques have limitations. Many times, this is due to the lack of involvement of end users of the data in the process of insight creation. The developers after consultation with middle layer management, apply transformation rules using data integration & analysis tools.

Developers interpret the business requirements as per their understanding and write the transformation logic. Several times, the end result is erroneous and requires intervention to 'fix' the logic in an ad-hoc way as rebuilding the models from scratch would require a lot of time and money – to reiterate the point, this is where dataOps comes into play. Figure 3 depicts data transformation steps as a part of the model, test and evaluate process.

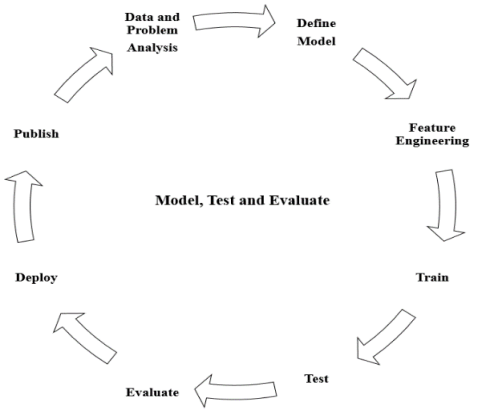

Figure 3: Data Transformation Steps

3.6 Transformation Strategies

Data should be transformed in such a way that it can be re-used for multiple scenarios. DataOps strategies can be applied during the lifecycle of the transformation process. Using agile & devOps methodologies, developers can build the transformation logic in an iterative way, and minimize the risks of loss of effort. The analytics need to be built in a rapid response fashion as well. The sales, marketing and other business teams need to be involved in a way that allows quick iterations to be useful in delivering minimum viable products (MVP) in agile parlance.

As data goes through the data analytics pipeline, it should be put through the right filters and tests to ensure that data is trustworthy and its quality is unquestionable. This is also where standardization brought in by dataOps plays an important role. Traceability and continuous flow of value are key in ensuring business value delivery using dataOps.

3.6.1 Identify key stakeholders

Identification of key stakeholders during transformation is the first step of the process. Some of the key stakeholders are the end users, business analysts, data analysts, developers, data scientists, architects and data modelers.

3.6.2 Team formation

A dataOps team needs to be formed with representation from identified set of stakeholders. This team needs to work in a collaborative manner beginning with definition of objectives.

3.6.3 Define objectives, roles & responsibilities

Objectives need to be clearly defined by getting the views from every participant. Brainstorming workshop, Design Thinking and Delphi techniques can be applied to gather inputs from team. This is a repeatable exercise till the objective is achieved.

3.6.4 Development, build, test and review

Developers prepare for coding for the first iteration in the development. The result of transformation is reviewed by right set of stakeholders such as data analysts, data scientists and architects & so on depending on the roles assigned in a specific organization. As new set of data is captured, it can follow the cycle

3.7 Data Repository Management

3.7.1 Current Challenges

A good repository which can scale and also provisions for security and access control is the keys to success of data analytics. However, this is one of the big challenge business enterprises face today. With no clear objectives set, managers find it difficult to get a go-ahead from the business units. Therefore, one can find data repositories spread like mushrooms across lines of business in a large enterprise. Eventually, it becomes quite challenging to bring everyone into a single federated system.

3.7.2 Strategies

3.7.2.1 Determine Size

Right strategy for selection of data repository is very important; and a key input is the data volume. For any large business enterprise the data volume may be huge and expected to increase exponentially. Deciding on a proper repository would need a careful analysis of the requirements, current volume and future growth. Storage, Security

administrators and architects should decide this strategy. Most importantly, the scalability aspect needs to be given highest priority. As the data volume grows, the storage should expand seamlessly without the pangs of feeling elastic limits of the repository.

3.7.2.2 Formulate storage/security architecture

For any large manufacturing & utilities enterprise, there will be data from various functional units which will require security. The data pipeline itself will need security features for managing roles & responsibilities as well as for data transfers.

As data gathering is a continuous process, the storage and accessibility requirements may also keep changing with time. Business use cases will also continuously evolve with market dynamics. This requires a scalable architecture that can sustain the changes required for various business use cases.

3.7.2.3 Determine hardware, storage & software requirements

To begin with, a small environment can be built which can expand with time, based on need, to avoid any risk related to building a large initial setup.

3.7.2.4 Build Environment

Build scripts can be configured to shut down and restart the environment so that the environment management can be automated, including tasks related to environment maintenance and recovery.

3.7.2.5 Monitoring and Planning

Continuous monitoring and planning to improvise are the key stages of devOps as also of dataOps. With new business requirements, ability to on-board new data sources, transformation, access management, and security & so on are the aspects which can be monitored to aid planning. Continuous improvement of dataOps setup on a regular basis for success. Figure 4 depicts a monitoring iteration with a focus on storage volume scaling using DevOps.

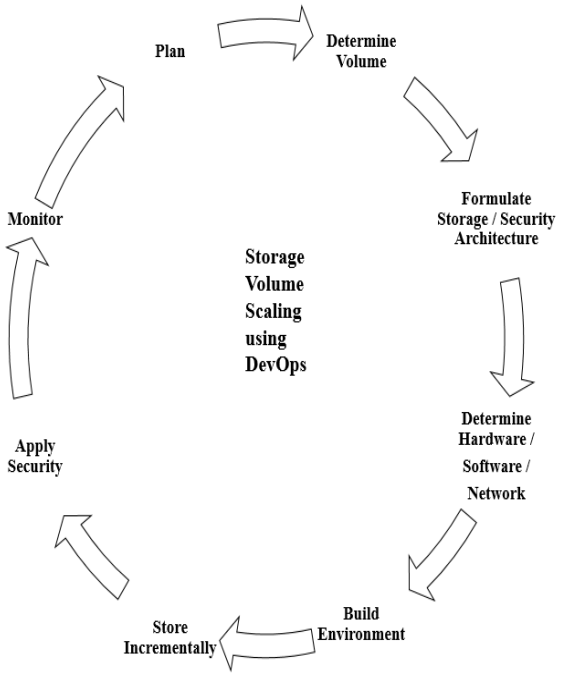

Figure 4: Monitoring Iteration

3.8 Discover

Discovery is a critical stage in this process cycle because at an enterprise level a number of use cases are solved in silos. Each time a problem is solved, the details are not necessarily available to other departments and therefore a lot of time is spent in potentially 're-solving' challenges. In a manufacturing or utilities enterprise, there should be a single model repository accessible across the divisions, departments. The question remains, how can one use this data to solve a problem which may not be known to others?

That is why before any problem is solved using a new model, the existing models and related data can be re-discovered if similar problem and solutions are available. The discovery process must allow for steps such as locating the owner & department and steps for enablement of reusability. This process not only minimizes rework but also helps optimize solutions with minimal possible effort.

3.9 Data Modeling

3.9.1 Current Challenges

Data models generally solve a specific use case, and re-usability is likely to be low. When an analytics model is developed, focus is on the problem at hand, not on reusability or growth. Generally, customer data is gathered in a text file, extracted and a model is built and experimented upon. A team comprising of data scientist and developers picks a customer problem and data. The data scientist would analyze the data and select the feature and algorithm. The developer would code and build an ad-hoc solution or a quick prototype.

This prototype would be based on small set of data that has been shared by a specific customer for a specific aim which defeats the larger purpose of dataOps in later stages. The code never gets re-used as it is not close to the reality of solving any actual problem. This is why a scientific approach to dataOps adoption is required which can aid sustenance of the model, encourage re-use, help growth and enable insight creation to help business.

3.9.2 Define data model

Data model plays a key role in AI/ML (Artificial Intelligence / Machine learning) area. From the repository, data can be selectively chosen to build a model. Different use cases would need different models. Creating a single model for all use cases would complicate the data model. For example, delayed computation of orders in a transport use case may require features such as delivery window, type of material, quantity, delivery unit location and so on which may be based on the past data patterns but this may differ from the model focusing on customer demand.

3.9.3 Train, Test and Evaluate

Machine learning methodologies can be applied on data. Depending on the requirement right methodologies such as supervised or unsupervised learning techniques or semi-supervised learning methods can be applied on the data.

Though there are established models & methodologies available, creating a proper trained model is always a big challenge. Established techniques should be readily available within the organization and they should allow for reuse. Problem solving in isolation should be discouraged which would be time consuming and counter-productive.

3.10 Insight Publications

Standardized machine learning methodologies can be applied on data to create models. These models need to be published as operational models at an enterprise level to aid business development. The details about the models need to be indexed as meta-data. This would help during enterprise search operation and in reusing already operational models with simple modifications rather than re-inventing the wheel.

3.11 Insight Consumption

The published models can be reused at an enterprise level to derive various analytics required for business. For example, "Recommend" solutions for understanding consumer preferences can be applied to a host of product lines. The tested and deployed solutions can be used with similar sets of data to solve similar problems. This way not only the enterprise saves time in quickly applying proven methodologies but also ensures building of robust solutions by continuous evolution process. Figure 5 shows a model, test and evaluate iteration.

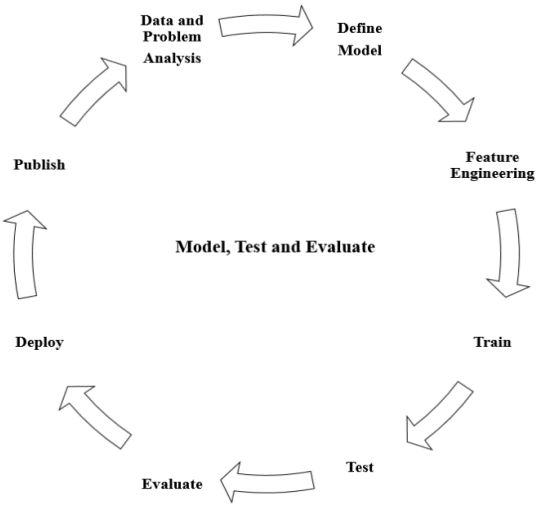

Figure 5: Model, Test & Evaluation Cycle

3.12 Monitor, Archive and Plan

Continuously monitoring the steps provides scope for improving the errors or inefficiencies in each step. These identified challenges should go into the model backlog. Various metrics can be collected for monitoring the health of the dataOps system. These metrics will help to understand the consumers of various analytics solutions, their consumption patterns, access threats, performance of models and capacity planning and so on.

Continuous monitoring is also important to aid understanding, monitoring and control of data in a disciplined fashion. With time, data may need to be purged or archived. Proper strategy for archival and purging must be planned. Sometimes, the organization might decide to cull out some irrelevant data before storing or archiving to ensure size management as well. Unwanted data needs to be purged. Periodic data audit can be conducted to help streamline such strategy.

4. CONCLUSION

The challenges mentioned in this paper are barriers for building data analytics systems which are crucial to sustained success of any manufacturing & utilities business. The success of products such as connected devices and autonomous vehicles is heavily dependent on relevant and timely use of statistical and predictive analytics models. This is so because embedded devices require a high amount of computing & decision making power to make critical decisions driven by enormous amounts of data they generate.

Therefore, building a data eco-system based on dataOps which is resilient and scalable, high performing and highly available requires a scientific and disciplined approach. DataOps has high potential and when enabled in the right way, can bring a lot of business value to the adopting manufacturing & utilities organizations. DataOps cycle is repeatable and iterative with scope for continuous evaluation & improvement of models. This will be the backbone of data analytics systems in the years to come.

5. REFERENCES

[1] Lismont, J., Vanthienen, J., Baesens, B., Lemahieu, B.: Defining analytics maturity indicators: A survey approach. International Journal of Information Management 37(3), 114–124 (2017)

[2] Knabke, T., Olbrich, S.: Capabilities To Achieve Business Intelligence Agility – Research Model And Tentative Results (Research-in-progress). In: Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS), p. 35 (2016)

[3] Baars, H., Ereth, J.: From Data Warehouses to Analytical Atoms – The Internet of Things as a Centrifugal Force in Business Intelligence and Analytics. In 24th European Conference on Information Systems (ECIS) Istanbul, Turkey (2016)

[4] König, L., Steffens, A.: Towards a Quality Model for DevOps. In Continuous Software Engineering & Full-scale Software Engineering, p. 37 (2018)

[5] Mishra, A., Garbajosa, J., Wang, X., Bosch, J., Abrahamsson, P.: Future directions in Agile research: Alignment and divergence between research and practice. Journal of Software: Evolution and Process 29(6), p. e1884 (2017)

[6] Pinkel, C., Binnig, C., Haase, P., Martin, C., Sengupta, K., Trame, J.: How to best find a partner? An evaluation of editing approaches to construct R2RML mappings. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. eds. ESWC 2014. LNCS, vol. 8465, pp. 675---690. Springer, Heidelberg 2014

[7] Evelson, B., Kisker, H., Bennett, M., Christakis, S.: Benchmark your BI environment. Technical report, Forrester Research, Inc., October 2013

[8] Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., Walle, R.V.D.: RML: a generic language for integrated RDF mappings of heterogeneous data. In: LDOW 2014

[9] Knap, T., Kukhar, M., Machá , B., Škoda, P., Tomeš, J., Vojt, J.: UnifiedViews: an ETL framework for sustainable RDF data processing. In: ESWC Posters & Demos 2014

[10] Wynne, M., Hellesoy, A., Tooke, S.: The cucumber book: behaviour-driven development for testers and developers. Pragmatic Bookshelf (2017).

[11] Schwaber, K., Beedle, M.: Agile software development with Scrum. Prentice Hall, Upper Saddle River (2002).

[12] Beck, K., et al.: Manifesto for agile software development. http://agilemanifesto.org/, last accessed 2018/07/29 (2001).

[13] Zimmer, M., Baars, H., Kemper, H.-G.: The impact of agility requirements on business intelligence architectures. In: Proceedings of the 45th Hawaii International Conference on System Science (HICSS), pp. 4189–4198. IEEE (2012).

[14] Krawatzeck, R., Dinter, B.: Agile Business Intelligence: Collection and Classification of Agile Business Intelligence Actions by Means of a Catalog and a Selection Guide. Information Systems Management 32(3), 177–191 (2015).

[15] Larson, D., Chang, V.: A review and future direction of agile, business intelligence, analytics and data science. International Journal of Information Management 36(5), 700–710 (2016)