**Introducing the DataOps Methodology**

Hello and welcome to the DataOps methodology - how to unlock data for productive use. DataOps is defined by Gartner as a "collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and consumers across an organization. Much like devOps, DataOps is not a rigid dogma, but a principles-based practice influencing how data can be provided and updated to meet the need of the organization's data consumers. The DataOps methodology is designed to enable an organization to utilize a repeatable process to build and deploy analytics and data pipelines. By following data governance and model management practices, they can deliver high quality enterprise data to enable AI. Successful implementation of this methodology allows an organization to know, trust and use data to drive real business value. This course is intended for anyone and everyone who wants to learn about data pipelines - whether they want to just understand what the considerations are in specifying data requirements, for example, C-level executives, managers, sellers and students, or if they want to play a role in the data pipeline itself, whether that be a data steward, a quality engineer or a data engineer. So welcome to this course which is all about getting to know what data you have available to you, so that you can mine it for insights into how to drive business success. Now, there are many ways of getting data that can be used to support a hypothesis or to form an opinion. but without a definite rigour and thoroughness in the way that data is gathered and used, questions can rightly arise about the wisdom of acting on the conclusions presented. Would you make a decision to invest half of your salary on shares in a company based on the sales of one store in high season? The data you use needs to be trusted and complete, and in order to provide the reliability and comprehensiveness of data that is used as the raw material for decisions, we need a robust, repeatable process. And not only is the step taken to provide the data important, we also want that data fast! The purpose of this course is to share a methodology that can be used to provide a repeatable process to provide data to data consumers, whether they are data scientists or business analysts, or any other person in any organization looking to gain insight. A methodology allows us to ensure that the data used in problem solving is relevant, reliable and traceable to address a question at hand. It provides a set of methods, guidelines and practices, that, when used, improve probability of getting the desired outcomes. In this case, that the right and trustworthy data will be delivered to the people who want to use it. This course is built on the experience of many practitioners at IBM and addresses every action that needs to be taken from data collection to productive use. Using a methodology produces numerous economic benefits. The data is dependable because it has been brought through a standard process. The data is more commonly used, because it is shared in a catalog with characteristics that make it easy to find. How the data is used is also shared, so people can learn from each other and springboard from previous data work to new ideas, enabling data innovation that builds on previous insights, avoiding rework through ignorance. The combination of methodology and integration automation that is embodied in software tooling maximizes the speed to delivery of insights with reliability and minimizes the chances of operating without relevant knowledge.

## DataOps Methodology Phases

In a nutshell, the DataOps methodology consists of three major phases. Establish DataOps provides a guidance on how to set your organization up for success in managing data. It provides all of the aspects that will determine success in delivering data. The data that is required to make good, future proof decisions and build unbiased models typically comes from multiple data sources. If we want the data pipelines that we lay down to feed these data end-users, we want to invest strategically. Spending time on a data strategy ensures that we've thought through our desired future state, before we spend resources on data infrastructure. Work is done by people, and these people are typically organized into teams. This is traditionally an area where data governance efforts fail because of communication issues - either because the teams are siloed by business function, or more commonly because data is owned by IT and used by the business - and there's a huge chasm between these communities. DataOps address this by ensuring the work is seen as important through executive sponsorship, and by the formation of multi-disciplinary, cross-functional teams. Automation is a key enabler of delivering data quickly, in a repeatable way. Establishing a toolchain ensures that every learning step from previous iterations are put into play for this and subsequent iterations. A Toolchain integrates the tooling, instrumentation and workflow required to orchestrate the collaboration that will make or break the production of data, eliminating wastes and delays. DataOps embraces the benefits of data governance, and part of establishing DataOps is putting "just enough" control in place to enable each data iteration to add within that structure to add new data governance artifacts. Establish Baseline provides that framework to draw the guidelines, allowing each iteration to "color within the lines". Once the first-rate team has been established, they need to be able to take on projects continually, without wasting time debating what the next top priority is. Ensuring that there is a pipeline of well-defined and consistently scored data sprint's on a backlog ensures continuous delivery of business focused data pipelines, building on previous deliveries and satisfying the business need for data as defined by the priority. The second phase is Iterate and Iterate DataOps delivers the data for one defined data sprint. It describes the considerations that need to apply during the conveyor belt of data from source to catalog. The institutional knowledge of data sources is often lacking due to changes in personnel over time, due to acquiring data from unfamiliar organizations, or for other reasons. Even when the right people and documentation is available, this can be a slow process to understand data sources beyond the metadata that is available in system catalogs. Discover allows the data to describe itself, using automation to recognize the patterns that exist to infer semantics, for example, recognizing the data values as being credit card numbers. Classification allows data to be transformed into information - by adding context and a common lexicon. For example, by standardizing on a common glossary, we can provide clarity on which of our database tables that have a column called "Nm" are actually customer names, and which are employee names. Furthermore, we can build a semantic graph of characteristics of each domain (for example customer, product, location and so on) that unifies the information stored in multiple places - allowing the information to determine the data available, rather than being constrained by how it is organized within data repositories. Moving onto to quality. In an ideal world, we would only use perfect data to drive our decisions, but data is often gathered manually and processed by software that contains bugs. So this step identifies how to assess the quality, and if required to mitigate any poor quality. Simply knowing the quality of data is in itself a valuable piece of data, which can be used to temper the actions taken. Policies and rules define how we can use the data. Often driven by, and based on the legal and regulatory requirements of an operating environment, these are often supplemented by an organization's own approach and ethics, and must be taken into account, particularly where they control the use of and distribution of data. A key aspect of the use of data is manipulating it to suit how it will be consumed - so many thousand transactions may be summarized, for example, by relevant time period, or data from an internal ordering system may be combined with data from an external benchmarking provider. It is important to be able to both manipulate and document the processes utilized for repeatability and traceability. This may provide the data required or it may provide an environment in which to experiment to develop the "recipe" for data wrangling that is needed. Data Movement is used to either virtually or physically move a combination of data sources with associated transformations to another target. This may be because of restrictions in the usage of the source (such as not wanting to impact operational systems) or because of latency issues - we're using the right ETL tool will provide the data that's required with regularity to the consumer within their defined service delivery metrics. Throughout the cycle of process and activities needed to deliver data, there may be opportunities to address issues. Remediation consists of analyzing the issues, determining where the issues should be addressed, at source or at some point in the delivery pipeline, and providing a plan of action to address the issue. Finally, the third part of DataOps is to continually improve the process. Each iteration is an opportunity to learn where things could be done better, every data sprint will bring different challenges and these learning should be channeled back into the process and teams to ensure that there is a virtuous cycle of improvement.

## Course Structure and Learning Objectives

## This course is broken down into six modules, each providing the understanding needed for the next one. Module 1 and 2 are all about establishing DataOps, and these follow the steps that we have just covered, building on any existing data governance efforts. It's helpful to understand the many aspects of team organization, prioritization, tooling choices and alignment with overall organizational strategy that needs to take place to start a DataOps initiative and set it up for success. Having set up the principles of engagement, module 3, 4 and 5 describe what happens for delivery of a given data sprint, which focuses on a key business objective and is measured by a set of KPIs that determine its completion and success. These modules address knowing your data, trusting your data and using your data for a set of data tasks that deliver business value. Module 6 allows for a process of reflection and improvement of any part of the process - whether that is improving the ways in which teams are set up and collaborate, or whether the experience with the current governance artifacts would warrant examining alternatives to be used on the next iteration. Throughout the course, we will be working through a case study that supports what you've just learned. And we will also include a number of review questions to test what you've just learned. When you're ready, you can take the final exam. After completing this course, you will be able to: - understand how to establish a repeatable process that delivers rigor and repeatability; - articulate the business value of any data sprint by capturing the KPIs the sprint will deliver; understand how to enable the organization's business, development and operations to continuously design, deliver and validate new data demands; promote necessary cultural considerations, for successful implementation of a DataOps practice including patterns that can be shared between teams as templates to promote successful adoption across an organization; understand our framework that fosters collaboration between contributors to the data pipeline towards a common business goal.

## DataOps Case Study

## So let's talk about the case study that we'll use throughout the course to illustrate different topics of the methodology. We're going to focus on credit card fraud analysis. A 2018 pan-European survey by Visa found that people are moving away from cash payments including mobile transactions, online payments, and credit cards, with over 80 percent of all European citizens using their mobile phones, a trend that is reflected globally. According to the 2019 World Payments Report, the transaction volume of non-cash payments reflects this trend and is growing rapidly, particularly in developing markets, with Asia at 32% growth and CEEMEA (Central Europe, Middle East and Africa) at 19% growth. It's projected to top 1,046 billion non-cash transactions globally by 2022, which is a compounded annual growth rate of 14%. One of the down sides of this reliance on phone apps and "plastic money" is the loss of funds to fraudulent transactions. The US has the dubious honor of leading the ranking table for credit fraud, with 38.6% of reported credit fraud losses in 2018. So what can we do to mitigate the risk? We will look at a hypothetical set of data that we are going to use to predict, when a transaction is not legitimate. We will describe data sprint in terms of the overall business objective, and put forward a set of KPIs that the bank's DataOps team determine will allow analysis of non-cash transactions to detect fraud ideally infusing the intelligence that can be encapsulated in an AI model to detect and stop fraud at the point of time that the transaction is actually taking place. Note that the purpose of this case study is to illustrate the principles of finding, using and integrating data for the purposes of a DataOps project. We are not proposing to have the gold standard in fraud detection!

## Introduction to AI Ladder and DataOps Capabilities

## So let's talk about what it takes to create a trusted foundation that is ready for AI and indeed for any productive use of data to fuel reliable decision-making in an organization. How can organizations unlock the value of data available to them? We see it as a series of foundational steps that are needed to help accelerate the journey to AI and that can be done by building a trusted data foundation that's ready for analytics. Let's walk backwards from the end goal of using data well for every business decision. If we think about wanting to infuse the ability to make the right decision on an ongoing basis should I sign this customer up for auto insurance, and if so, at what premium? we need the ability to analyze the data available to us from all relevant sources, ensuring we have thought of every possible contribution to making that decision. Being able to get to the right data, means that the data needs to be organized for use understood, checked for quality and prepared in a way that we can start to analyze that. In short, we need to be able to know what the data means, we need to be able to trust it is accurate, and we need to be able to combine it from multiple places to use it. The first step is identifying the different data sources that we want to make available to people, giving them the right access - regardless of where that data lives. And we're not just talking about traditional data sources, that are typically in relational databases, so very structured, but also, about new data sources, such as emails, blogs, webpage content and so on, these can reside in the company's data centers across globally dispersed regions, as well as external sources. Organize is a step where we curate and share the knowledge of that data. We answer questions such as what does the data mean, where did it come from, can I trust it, can I measure its quality. In the same way that organizing my own things at home will certainly help me, but is crucial if I want anyone else to be able to find it. So in an organization, improving the findability and understanding about a data asset makes it easier and safer to reuse and share with others. Once the data is readily understood and made available, we can start to use it at scale, channeling it into analytics and AI models where we can derive insights to drive the business forward for success. Once we can create this data and the analytics pipeline, we can infuse this knowledge based insight throughout the organization, making better decisions based on real quality data. The focus of this course is on organizing the data through a repeatable process, or in other words, how do we make sure the pipeline works every time and that the data continues to flow to where it is needed. While the focus of this course is not on the software tooling that supports the methodology, it is worth spending a few minutes on how the methodology needs to be supported. The success or failure of a DataOps engagement relies heavily on end-to-end integration and on the use of automation to deliver results quickly. The path to Organize data starts with bringing the data into a well-defined process that starts with discovery (of dark and unknown data) or ingest (through well-defined data ingest pipelines). Automation is a key part of this process to be able to truly scale to meet the volume and variety of data coming in. Data brought into the governance pipeline has to be "understood", in other words, ensure the technical data makes sense from a business point of view. For example, you might find data of the form 123-45-6789 and you can guess that that is a social security number or an employee ID or both. The classification and business term assignment step adds business meaning to this technical data. Automation is really important here, since classifying data is often one of the most significant challenges data curators face, as data volumes grow. Then comes the curation and workflows, quality assessment and controls over this data to ensure the data is "good". Assessing quality is really important to establishing trust in the data being used, and mastering data ensures that we understand the people and organizations we are interacting with, no matter what touch points the data we've collected has come from. Data virtualization and data movement ensures that data is available when needed, just like water flowing in our taps, with no perceptible delay. Finally the cataloging step that makes this data available across the enterprise ensures that data consumers can take data into their own hands and start to use it to answer questions whether that's for building models, for testing applications, or for driving business decisions. Returning to one of the key objectives of delivering business ready data quickly what slows that delivery time? Data pipelines are the primary source of bottlenecks. When an executive makes targets like "I want to increase sales revenue by 5% next quarter"; "I want to decrease cost by 5% over the next three quarters"... the answer is data hungry and data sensitive. The team reporting to her may look at multiple options that will determine the measurement of the business goal and decide to go with a couple. They need data to prove and to execute the targeted action. They involve the Business Analytics and/or the Data Science team to use data to drive actions. With so many data sources available, the team needs to figure out where the required data is located. Then they need to identify the right tables and fields - especially custom fields, to choose. Let's think of this from the mind of a data scientist who's helping to deliver the data. They'll ask questions like: "which revenue field is relevant? There are 7 of them, including three custom fields! Ok, let me pick this one ... I had heard someone using this field in one of the meetings". They face data that's not complete and not clean. They struggle with needing to remove duplicates, to clean values. They may ask - how many varieties of phone number, zip code, account names exist? The problem here is that data pipelines are not open and clear to the data scientists - and indeed the team, who are trying to get the data to the executive. The data pipeline is the bottleneck to deliver data to users. We need to figure out how to reduce the time to deliver that data without compromising quality. When you look at what customers are experiencing today, you'll find that it's taking months and quarters for their organizations to conduct basic data operations to discover what data they have to understand it, to ingest it, to integrate it, to assess its quality and to remediate the data quality. This is what DataOps is here to solve.

## Data Strategy Overview

## Let's look at how we can use a data strategy to understand our current data landscape and evolve to where we want to be. Forming a strategy for data is one of the first things that you need to establish when preparing for a DataOps approach. It examines your current data landscape and prioritizes what the organization wants to maintain, and what it wants to move towards changing, and provides a roadmap from the organization's current set of data repositories and their approach to managing data and where they want to be. A data strategy connects your business plan and priorities to your data, AI and analytics requirements. A data strategy is a foundation for becoming a data-centric organization. An effective strategy recognizes data as a strategic asset with its value being demonstrated across multiple use cases, initiatives and lines of business. It drives innovation through new projects and initiatives. It identifies and removes blocking factors from existing projects. It includes an overall architecture and it is actionable. Outputs of a data strategy are a roadmap and action plan. These are developed in consideration of your current state to drive innovation and remove blocking factors. So a data strategy is a plan of action that can be put in place to deliver the data needed to support an organization's business goals. It helps to clarify both the nature of data required, and how that data needs to be managed. An organization's data strategy is embodied by their data architecture, which describes how data is collected, stored, transformed, moved and consumed. A data strategy needs to be clearly articulated before you start to engage on DataOps. Why is that? Because it provides a framework for deciding how to manage data throughout the organization. For example, the data strategy defines how to determine data quality, where and when it is appropriate to tolerate differing levels of quality, and when it is necessary to undertake remediation for data quality. A data strategy can be seen from two perspectives: what needs to be done from a mandatory perspective, and what data strategy will put the organization ahead of its competitors. The execution of the strategy guides the organization by a series of tactical steps towards the overall architectural goal. It includes the rules governing data repositories, such as databases and file systems, and the systems for connecting data with the business processes that consume it. Let's look at what a data strategy needs to include.hu

## Data Strategy Considerations

## What considerations are included in a data strategy? In the rest of this lesson, we're going to focus on aspects that need to be considered when addressing the data needs of an organization, including users and consumers of the data, flows of the data and constraints including regulatory requirements and risk management. For example, for regulatory requirements, what constraints, such as data privacy rules, do we need to bear in mind? For risk management, what information do we need to capture to provide an understanding of the possible damage that could occur to our business? We need to consider data origin or sources - where can we obtain information from? What is our current data topology and are we seeing any operational issues that we need to address? Are we happy to only look for data internally, or are we open to including data from external sources, such as benchmarking data? We need to consider data monetization - how can we gain measurable financial advantage from the data that's at our disposal? We can look at this in terms of cost savings, reducing what we have spent, or as a net gain that takes advantage of our information assets to drive revenue for our organization. We need to consider skills management where are we strong in our knowledge of how to manage data, and is it strategically relevant? Do we need to invest in new skills to equip our employees for future data management and usage? We need to consider data quality while we may always strive for "perfect data", how much signal noise due to dirty data are we actually willing to put up with - and for what sort of data? We need to consider expense management every decision to change and improve has an associated cost, in infrastructure, in potential lost opportunities when we do nothing, and in investing in our skills. How much do we want to spend, and what's the potential return on that spend? All of these aspects need to inform our data and AI architecture and action plan in order to meet the business needs. A data strategy should include data management decisions determining the organization's actions and in respective to regulatory requirements how regulated is the industry in which the organization is operating? What's the importance of fulfilling the requirements of regulation relative to the cost of compliance? For example, if by implementing a program, an organization might spend ten times as much in paying any fines that might ensue, and if the likelihood of such payment was low, a decision might be made to make tactical efforts for compliance. On the other hand, if there were severe consequences of non-compliance both from a financial impact including loss of business and fines, as well as from a reputational standpoint a decision might be made to heavily invest in ensuring high levels of compliance. In general, most applicable regulations leave an organization with little choice but to comply although the regulatory burden is considerably higher in some industries, such as health care and financial services, which reflects the wider societal impact, where the regulation is not in place. Gartner defines Integrated Risk Management as a "set of practices and processes supported by a risk-aware culture and enabling technologies, that improves decision-making and performance through an integrated view of how well an organization manages its unique set of risks". A risk can be quantified as the estimated possibility that a future event will prevent an organization from reaching a goal, and that risk can come from a wide variety of sources whether caused by failure of the organization itself to meet operational goals, or external events such as stock market instability. Management of risks depends on understanding key indicators, being able to baseline an organization's position relative to that risk, and evaluating how that position changes over time all of which require consistent and reliable sources of data. As technology and business practices evolve and change constantly, a data strategy needs to consider what skills the workforce of the future needs to have. Traditionally data strategies focused primarily on the skills that were needed to run the data centric functions of the organization - the "IT shop". As companies move to more of a self-service model of data usage, the data strategy needs to focus on multiple types of data providers, curators and consumers, and to plan for the skills required to fully take advantage of their information architecture. Where is the organization comfortable with sourcing its data? What is the strategy for addressing systems of information that are showing signs of age and embodied practices that are no longer relevant. For example, several systems in one financial institution included database tables where the first and last columns were never used. When investigated, the reason behind this practice was that: "Punch cards tend to fray at the edges"! More generally, in many cases, organizations have simply had to make do with the data available to them internally if they had accurate information on their competitors business results, they might be able to drive greater competitive impact. The emergence of data brokers opens up the possibility of being better informed on market dynamics, assuming that an organization's data strategy includes an ability to incorporate external data. What is the fault tolerance for the use of of data in the organization? Any source of data that involves manual data collection will inevitably result in data quality issues whether that is the presence of inconsistent values, missing data, duplicates, or other issues that make the interpretation of information difficult or unreliable. Data strategy determines the key uses of data where high-quality data is of paramount importance (such as financial reporting) and areas where there might be a higher tolerance (such as determining new markets where the investment is not substantial). Technology is constantly evolving, and many potentially trans-formative and disruptive technologies, such as artificial intelligence, machine learning and natural language processing promise huge benefits saving time and improving efficiency by automation, freeing up staff to work on other or cognitive intensive areas. But every spend represents an allocation of the organization's capital that could potentially be better used in other areas of the business. The data strategy needs to include a detailed cost benefit analysis of any plan to support the use of data in the organization. On the other side of the balance sheet is the potential to drive incremental revenue based on wise investment. Data strategies that drive the ability to cash in on the investment they're making in their data. The financial benefit can be on minimizing cost in a company's operations, decreasing waste and increasing productivity, maximizing opportunities for their products and services by continual improvement of what they offer and by providing the best of class customer experience. It can also include the creation of new revenue streams by making data available to customers and partners. Both are possible but rarely occur without focus and drive. These are two sides of the same coin and are impacted by an organization's attitude to spend. Over a long period of time, the number of data systems in use tend to grow, they are often entrenched in the way business is transacted, and a focus on their removal or in combining multiple systems even where business practices would suggest a benefit - gets put on a long finger. These older data systems are often bound to legacy and inflexible technology, and failure to address the problems with their continued use can have a significant impact, not only from the financial implication of software and hardware costs, but in inefficiencies of maintaining them. We have looked at a number of considerations in creating a viable and effective data strategy for an organization, from deciding the parameters on data origin, data quality and risk. Each of these factors that can change the cost and reliability of data and the decisions that are driven from that data, as well as the financial effect of data for cost reduction. Regulatory compliance is often seen as an obligation, and brings both financial and reputational consequences for non-compliance. A well-designed data strategy will consider all elements and design a balanced approach but no matter how good a strategy, it will require continuous adjustments, as the economic environment changes around it. By starting with an unconstrained business need, we can consider overall business objectives and outcomes desired, and build a roadmap to bridge the gaps to develop a data strategy that is more resilient to future needs, and lowers ongoing operating and maintenance costs.

## Data Strategy Case Study

## So let's switch to the case study that we introduced in our earlier lesson. Imagine that you are a chief data officer for this bank, and you have been alerted to the issue that the bank is hemorrhaging funds because of online fraud. You have recently joined, and are keen to demonstrate the benefits of automation that you consider could be employed through developing machine learning models, which can be deployed within the transaction verification process. When you investigate the data landscape, you find that there are several systems that seem to have the same sort of data and that needs to be addressed. Historically, because the bank has grown by mergers and acquisition, and so it's inherited all the technological infrastructure of all of the banks that have been in operation before they joined the larger bank group. What are the threats to being able to provide good quality data that you can imagine? A data strategy will prioritize this within an action plan, considering whether the bank should plan to consolidate multiple data systems into one with consequent impact on the operational systems that use each silo of data or whether to employ another data tactics, such as data virtualization which can make the multiple data silos act as one. Either way, the data strategy needs to balance the costs in any new target data architecture with the latency of accessing multiple data stores in place, and consider the use of the data, with real-time response times required for transaction processing being a critical factor in the plan. What are the opportunities that someone like you, a relative newcomer at the bank, can bring to defining a data strategy? A data strategy should strive to make a plan that is objectively beneficial to the organization. As a newcomer, you have the opportunity of brokering a plan that considers how data across the organization can be leveraged, without the shackles of prior ownership of one part of the business. This is a cultural consideration, and could result in bringing together the key stakeholders to work towards common goal. DataOps puts into place common tools and metrics that can result in a consistent application of quality goals, and the use of common business terminology to provide a standard view of data throughout the organization.

## Establish Team Introduction

## The DataOps methodology is designed to enable an organization to utilize a repeatable process to build and deploy analytics and data pipelines. The delivery of data using such a well-defined repeatable process is carried out by people, organized into teams. We've all had experience of teams that work well, and those that don't, and while there are many theories on how to reach the ideal team structure and dynamics, a key aspect of DataOps is focus on the business objective and rapid delivery of the data required. For this to happen, the team needs to be as focused as possible on the outcome, needs to include key skills from the business to understand what is required and how it will be used, as well as technical know-how to understand how and where to find the data required and what is involved in delivering the data pipeline for a particular business use case. In this lesson, we will meet the people that will typically make up a team. We will discuss who the stakeholders may include, and suggest some ways in which clear articulation of the goal and accelerated path to delivery can be made enhanced by the way a team is organized and how communications are delivered. This lesson looks at how a team can be formed bringing together team members with skills and different aspects that will form a cohesive set of skills to focus on the data required by the business. A key advantage of using a DataOps approach to delivery of data is increased speed of delivery. DataOps supports highly productive teams with automation technology to help deliver efficiency gains in both project outputs and time to deliver. However, to experience benefits, the internal culture needs to evolve to truly be data-driven. This desired speed of delivery can only be realized when there is a culture of communication and collaboration between all parties who are working together to provide the right data to the right data consumers. While this may seem obvious, historically, organization have been very siloed in the way that they approach data. And they're probably one of the last parts of the organization to consider moving from a waterfall approach to managing projects. Generally, data is "owned" by IT, which may be a central function in the organization, but it is managed for separate lines of business so that marketing systems that monitor client interest and provide a gateway to inform the customer, may be totally disconnected to the product managers who make decisions on what to offer the customer, or the relationship manager who ensures that the customer receives a positive experience. As organizations increasingly turn to a DataOps approach to deliver business ready data, it is important these organizations learn how to structure their teams for success. A key tenet of DataOps is improving the lines of communication so that requirements for management and delivery of data are clearly understood, and facilitate a rapid turnaround of focus delivery within the guidelines of the organization's data strategy. A good combination of a DataOps team thus combines both business knowledge and technical know-how. Teams are organized so that they have just enough people to deliver effectively with clear understanding of what they need to deliver every day towards their goal. A good team will include some members who have good institutional knowledge of the organization how it really works, who they might need to reach out to, how to get things done with enough new knowledge of how things need to change to embrace the DevOps culture when to push against resistance, how to get things done differently. DataOps relies heavily on the use of automation to operate its speed teams can use a variety of tools to distribute and organize workloads it's important that they standardize to one way in which the various tasks are orchestrated with the ability to calibrate at any point of time where they are in relation to their goal. And the value of data is not confined to the output of DataOps team the use of data to track progress or the Key Performance Indicators, referred to as KPIs, for a data sprint, is updated regularly, as each team member contributes towards the goal of the sprint, enabling the team and all interested parties to understand how close they are to being complete at any point in time.

## Team Roles and Responsibility

## Let's spend a little time introducing the people that will work on a DataOps team. Knowing the skills and background that are needed will allow you to assemble the right team. We want to build this team out of the ideal mix of abilities recognizing that, in reality, it can often be limited by the pool of available resources. Knowing the skills required can also help an organization to plan for the needs they will be able to meet in the future and feeds into and supports the data strategy they have in place. Overall, whatever a team member is responsible for on a DataOps team, they need to be good communicators. As each team member gets to grips with the data challenge involved, they will learn the questions to ask each other at the very start and throughout the project to ensure the objective is clearly understood. Frequent touch points will help keep everyone track for their mission, and KPIs will help them understand what they have achieved, and what is left to complete. Let's have a look at some of the active members of any DataOps team. Gartner defines a chief data officer as "a senior executive who bears responsibility for the firm's enterprise-wide data and information strategy, governance, control, policy development, and effective exploitation." The CDO's role will combine accountability and responsibility for information protection and privacy, information governance, data quality and data lifecycle management, along with the exploitation of data assets to create business value. The Chief Data Officer is the primary advocate for DataOps, which she sees as a way of being able to balance all of the objectives she has in managing enterprise data assets and driving the business forward, making decisions at every level based on facts. As Chief Data Officer, Sophia is responsible for creating an enterprise-wide data governance structure to manage data as an asset. She ensures that users have access to timely and reliable data. To deliver on her data goals, Sofia needs to establish the teams needed to deliver data according to the data requirements communicated by business sponsors, addressing any data issues. In order to meet the requirements of the pervasive use of reliable data, it is in Sophia's interest to pursue the democratization of data, making it easy to find and use. She needs to establish access to data that respects regulations, and enables data consumers to combine multiple sources of data easily. To establish trust in the data, she needs to ensure that quality management is a focus in the delivery of all data to consumers. A Data Steward is a key role in the task of curating the data in an organization, especially for data quality and definition, who works with the CDO office to develop a data governance framework. In an environment of many different data sources, Peter's job is to provide a semantic layer which simplifies the data landscape. This layer needs to be accessible to the data consumers, who are generally business oriented, so it needs to be able to translate from what the data means (the information content) to where the data can be found (the technical storage characteristics). The Data Steward understands the overlap and interplay of data in multiple systems and oversees the creation of a business glossary to provide common understanding of data. Creating business terms, and mapping out how the information elements relate to each other in a business sense is a starting point for Peter. This is a task that is without bounds. So within the DataOps practice, the definition of a business goal and KPIs can help Peter stay on track for a given data sprint. Within each iteration, Peter can expand the business glossary content, creating and maintaining standards for clarity in business terms, adhering to the vision of the overall data governance framework to provide meaning to data assets, and reporting and monitoring the quality of all cataloged data. Understanding data quality begins with a policy that determines the quality that an organization is striving for, and it is Quan's job to bridge the gap between this target and the tools used to implement data quality assurance. Employing repeatable techniques, such as quality rules, Quan improves the overall data quality and reliability through automating the detection, and, where possible, the mitigation of quality in the data pipeline. Quan is responsible for understanding the cause of data quality issues, determining the best approach to eliminating those issues, and ensuring that the CDO is aware of the level of quality in different elements of the information landscape. Quan helps data consumers by ensuring they are aware of the quality of all data assets being made available through the enterprise catalog. Robert is responsible for creating data pipelines that feed the data required from source systems to where they need to be delivered. As such, he is familiar with the multiple source systems in use, and endeavors to build reusable data movement components that can be readily understood by members of his team. His skills are in demand, as he and his team are often the bottleneck that slows down the data delivery. Robert relies on well articulated requirements for data, and often needs to spend a lot of time investigating the data to ensure that what is communicated is actually delivered. Quality impacts Roberts pipelines, because unexpected values can break the transformations that are employed. Robert's hope is to be able to build a resilient pipeline, which once laid down, can operate without further adjustment. Every element in defining the data requirements and curating data sources are a factor in meeting this goal. Kathy is someone who makes value out of data. They are known as a story teller with data. She proactively gathers information from various sources and analyzes it for better understanding about how the business performs and builds AI tools that automate certain processes within the company. Kathy acts on resolving challenges for the business by analyzing data and builds automation into business processes by building and deploying machine learning models. Kathy is concerned about making the right insights which are only as good as a quality and extent of the data she uses. Furthermore she needs to ensure that the automation already in production continues to perform well as the operational data changes with the business environment.

## Team Organizational Structure

## Each of the people we discussed until now work together in small, focused groups, coordinating the work through workflow tooling, updating the KPIs for the data sprint, and each contributing towards the focused goal of delivering business value. What makes a "small team" - typically between 5 and 10 people, each experts in their role. There is a flow of data from its raw form to being understood and measured for quality, before it can be assessed as being ready for consumption. Each step of the way uses as much automation as possible, and creates governance artifacts that will improve the next iteration. For example, new business terms may be created if the business use case involves a new area of the business. These are created with reusability and enterprise focus in mind, for reuse in future projects. Likewise, quality rules are created to find any issues with the data that will become part of the pipeline for all future data flows. The team uses dashboards to understand their progress towards delivering on the key performance indicators for their data sprint. For a team to work effectively, and for the overall DataOps initiative to function well, clarity of authority and expected function of each person is required. A RACI matrix (Responsible, Actionable, Consulted and Informed) defines the roles and responsibilities for processes or projects, to ensure that key stakeholders are involved in decisions and are held accountable. Many organizations are hindered by the lack of communication and clear responsibility when managing business workflows that support the authoring and maintenance of their data governance programs, which can contribute to costly delays and missed deadlines. Implementing a RACI matrix, and integrating those allocations within the governance tooling, allows for collaboration with other users in the RACI matrix while also providing transparency on the progress of each person's activity in real time. This ensures corporate accountability in the management of governance artifacts, such as business terms, policies, rules, reference data, data classes and classifications for your business. In this lesson, we've looked at the team structure that works well to deliver business-ready data quickly for DataOps. But how does that team fit within the whole organization, and how does it understand its role within the organization's initiatives? Let's describe a typical organizational structure that supports the work of the DataOps team. The actual group structure and reporting lines of the individuals working on DataOps can vary, what's crucial is the recognition of the importance of the enterprise-wide effort to provide good and reliable data. Any initiative that is to succeed must have an executive sponsor, who confirms the importance of data to the organization. This can be any executive with the power to allocate budget and resources and influence the rest of the C-Suite, although it's most commonly the Chief Data Officer. The executive steering committee is made up of executives such as the Chief Information Officer, Chief Risk Officer, CDO and line of business executives, and is responsible for the actual selection of initiatives and allocation of budgets accordingly and has an important role to play in providing resources to the DataOps initiatives.

## The Enterprise Data Governance Council includes the data governance manager, business process owners, and compliance and legal expertise - who determine the policies and rules on data usage and provide guidance on appointees for RACI purposes. This team resolves any disputes on any policy issues. The data architecture working group includes enterprise data architects, data modelers and database administrators. This is a team that the DataOps team works with to address any issues with how data is stored, and plays a key role in fulfilling the data strategy. The Data Governance Office includes a metadata administrator and data governance analysts, who carry out the execution of policies set out by the Enterprise Data Governance Council. The DataOps team can be, but is often not a separately resourced team to the individuals shown here contributing to the well-being of data throughout the enterprise in different ways. Rather the function of each team contributes to an overall goal that is driven by speed of delivery for the business goals set down by the Executive Steering Committee. It is this fertilizing of the DataOps team by expertise from all aspects business, data semantics and structure, and technical know-how, that is the key to the success of DataOps. In this lesson, we've looked at the people aspect of DataOps. Skills and expertise are the lifeblood of a successful DataOps team the right subject matter expert, a skilled data steward and quality analyst with the knowledge of the organization's data systems and how they can be transformed into a data pipeline. and how they can be transformed into a data pipeline. The right mix of the right team members changes from project to project as the business use case changes. Once formed, the team can stay on track by continually measuring the KPIs defined for the data sprint. The team knows what their individual tasks are by being informed by the collaboration software being used as part of the toolchain and clarified by the use of RACI. The availability of access to the right team members will generally not happen without management support, so executive stakeholder sponsorship and interest is key.