Optional: Overview of the DataOps Methodology

Gartner defines DataOps 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. DataOps aims to create predictable delivery and change management of data, data models, and related artifacts. DataOps uses technology to automate data delivery with the appropriate levels of security, quality, and metadata to improve the use and value of data in a dynamic environment." (Source: https://blogs.gartner.com/nick-heudecker/hyping-dataops/) A small team working on a simpler or limited number of use cases can meet business requirements efficiently. As data pipelines and data infrastructures get more complex, and data teams and consumers grow in size, you need development processes and efficient collaboration between teams to govern the data and analytics lifecycle. From data ingestion and data processing to analytics and reporting, you need to reduce data defects, ensure shorter cycle times, and ensure 360-degree access to quality data for all stakeholders. DataOps helps you achieve this through metadata management, workflow and test automation, code repositories, collaboration tools, and orchestration to help manage complex tasks and workflows. Using the DataOps methodology ensures all activities occur in the right order the right security permissions. It helps set in a continual process that allows you to cut wastages, streamline steps, automate processes, increase throughput, and improve continually. Several DataOps Platforms are available in the market, some of the popular ones being IBM DataOps, Nexla, Switchboard, Streamsets, and Infoworks.
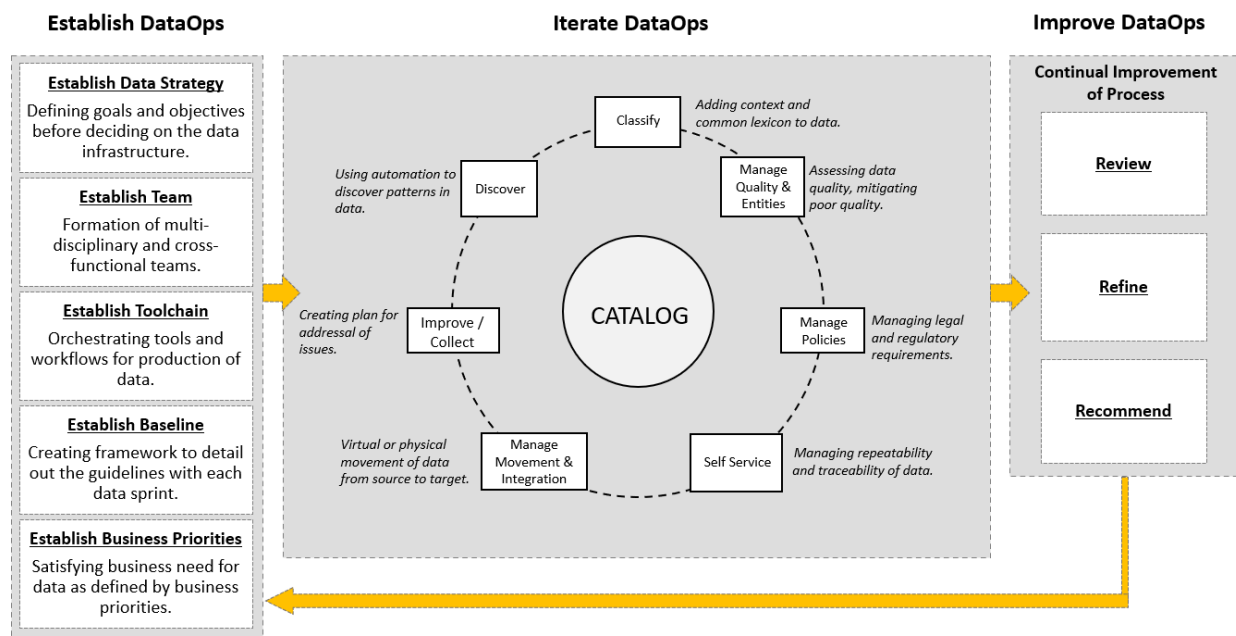
DataOps Methodology:

The purpose of the DataOps Methodology is to enable an organization to utilize a repeatable process to build and deploy analytics and data pipelines. Successful implementation of this methodology allows an organization to know, trust, and use data to drive value.

It ensures that the data used in problem-solving and decision making is relevant, reliable, and traceable and improves the probability of achieving desired business outcomes. And it does so by tackling the challenges associated with inefficiencies in accessing, preparing, integrating, and making data available.

In a nutshell, the DataOps Methodology consists of three main phases:

- The **Establish DataOps Phase** provides guidance on how to set up the organization for success in managing data.
- The **Iterate DataOps Phase** delivers the data for one defined sprint.
- The **Improve DataOps Phase** ensures learnings from each sprint is channeled back to continually improve the DataOps process.

The figure below presents a high-level overview of these phases and the key activities within each of these phases.

Establish DataOps

Iterate DataOps

Improve DataOps

Establish Data Strategy
Defining goals and objectives before deciding on the data infrastructure.

Establish Team
Formation of multi-disciplinary and cross-functional teams.

Establish Toolchain
Orchestrating tools and workflows for production of data.

Establish Baseline
Creating framework to detail out the guidelines with each data sprint.

Establish Business Priorities
Satisfying business need for data as defined by business priorities.

Adding context and common lexicon to data.

Classify

Using automation to discover patterns in data.

Discover

Manage Quality & Entities
Assessing data quality, mitigating poor quality.

CATALOG

Creating plan for addressal of issues.

Improve / Collect

Manage Policies
Managing legal and regulatory requirements.

Virtual or physical movement of data from source to target.

Manage Movement & Integration

Self Service
Managing repeatability and traceability of data.

Continual Improvement of Process

Review

Refine

Recommend

Benefits of using the DataOps methodology:

Adopting the DataOps methodology helps organizations to organize their data and make it more trusted and secure. Using the DataOps methodology, organizations can:

- Automate metadata management and catalog data assets, making them easy to access.
- Trace data lineage to establish its credibility and for compliance and audit purposes.
- Automate workflows and jobs in the data lifecycle to ensure data integrity, relevancy, and security.
- Streamline the workflow and processes to ensure data access and delivery needs can be met at optimal speed.
- Ensure a business-ready data pipeline that is always available for all data consumers and business stakeholders.
- Build a data-driven culture in the organization through automation, data quality, and governance.

As a data practitioner, using the methodology can help you reduce development time, cut wastages and duplication of effort, increase your productivity and throughput, and ensure that your actions produce the best possible quality of data.

With DataOps, data professionals, consumers, and stakeholders can collaborate more effectively towards the shared goal of creating valuable insights for business. While implementing the methodology will require systemic change, time, and resources, but in the end, it makes data and analytics more efficient and reliable.

Interestingly, it also opens up additional career opportunities for you as a data engineer. **DataOps Engineers** are technical professionals that focus on the development and deployment lifecycle rather than the product itself. And as you grow in experience, you can move into more specialist roles within DataOps, contributing to defining the data strategy, developing and deploying business processes, establishing performance metrics, and measuring performance.