Pluto7

DATA OPS

CONTENT

CONTEXT SETTING

Data is the stimulus for novelty and continuing a zealous advantage. It's the main factor for driving analytics, mastering business trends, and opportunities. Exploring the merit of data in novel ways can even expedite an organization's expedition to AI. According to Experian's 2019 Global Data Management Research report, 89% of businesses report that they struggle with managing data[1].

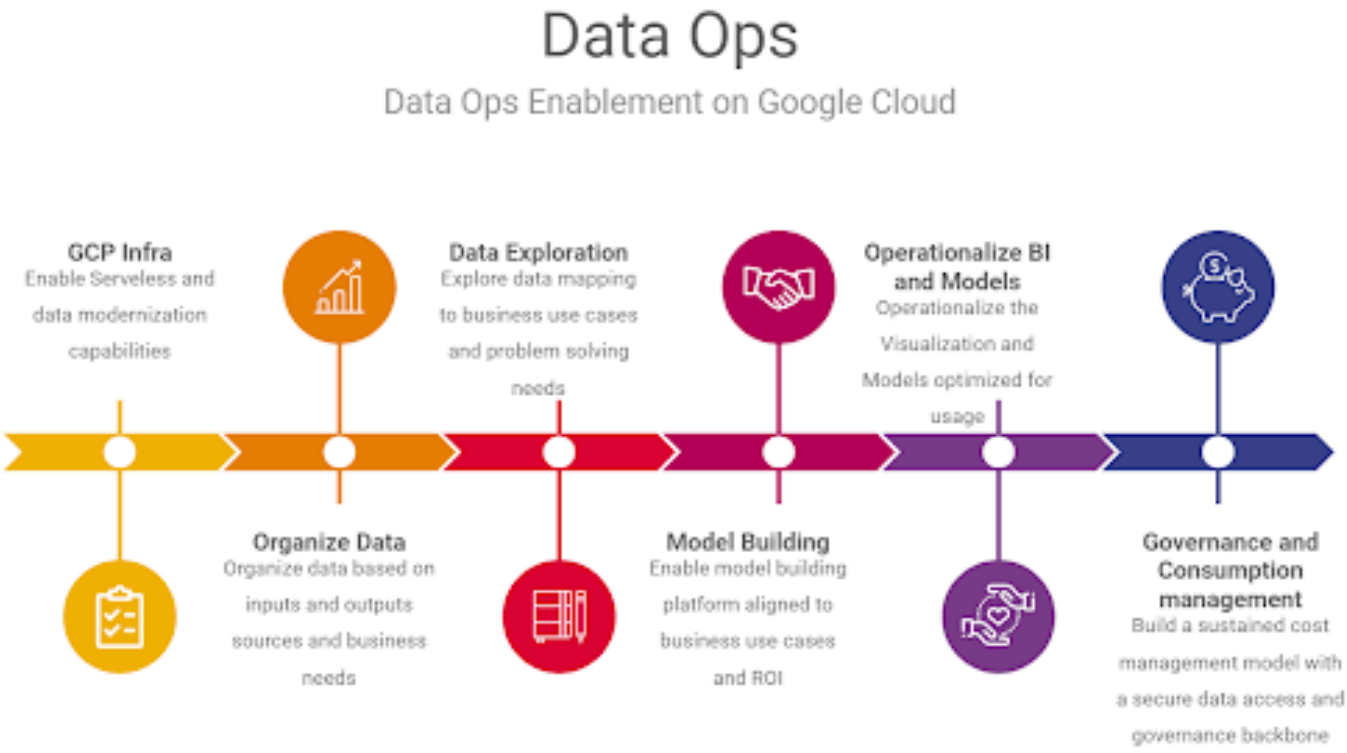
These tussles impede in understanding and insufficiency of confidence in underlying data. Understanding an organization's business goals is vital to evolve a productive data strategy for AI and analytics. Success depends on framing a process of data operations with a federated business ability to take up a data pipeline, which produces a complete and consonant view of the business all the time. Businesses all around are gazing for ways to upgrade their operational productivity and effectiveness to validate the optimal decision-making, mainly due to numerous silos within a company. These two factors influence business key persons to seek novel ways to address their main challenges within a framework.

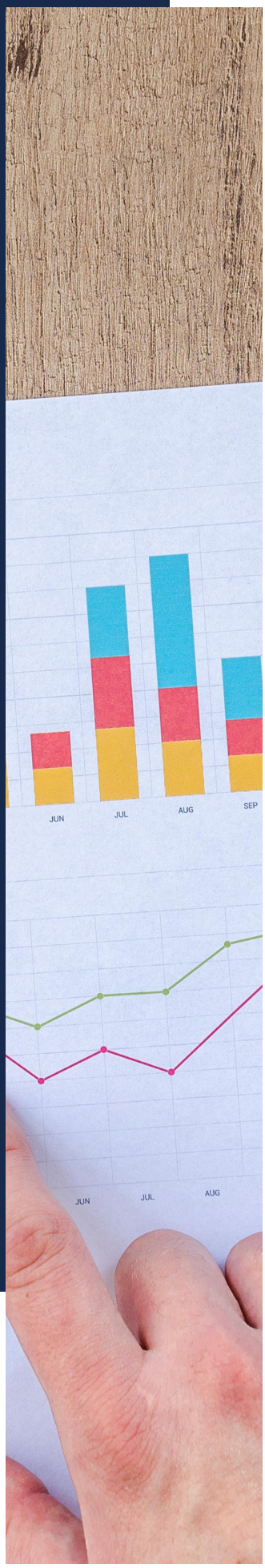
For companies seeking a revolution within their automation technology, data operations can implement a competitive advantage. Data becomes great when trusted business-ready data helps transform insights and operational accomplishment for companies. The goal of this white paper is to emphasize the benefits of the DataOps methodology, roadmap, and practice.

WHAT IS DATAOPS?

DataOps is a holistic approach for implementing data analytics solutions that use development, containerization, orchestration, testing, automation, collaborative, and continuous monitoring to continuously expedite output and improve quality to drive AI at scale. The purpose of DataOps is to expedite the creation of pipelines of data and analytics, automation of data workflows, and achieve high-quality data analytic solutions that meet business needs as swiftly as possible.

Here is a visual to show Dataops in a nutshell.

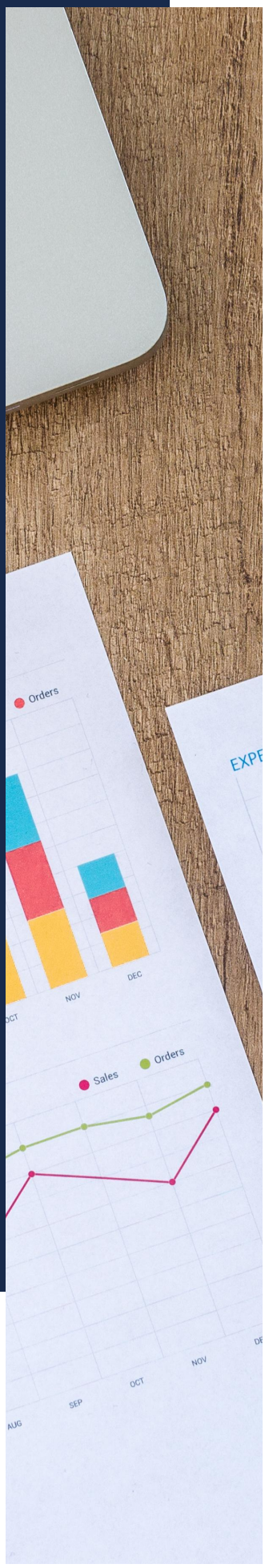




Fundamentals and Benefits of DataOps

Fundamentals of DataOps

- Data alone is not enough, it helps in interpreting the insights to deliver and add Business value to satisfy customer needs.
- Constant improvement can be achieved by understanding the process completely and continuously reviewing it, adapting to the improvements, and learning from the mistakes.
- By getting the updates of data analytic life cycles at every stage will help us in having a better understanding, collaboration, and communication.
- From process automation and reuse of pipelines, we can save time and manpower as we avoid rework and repeating the same steps for replicating/creating a new pipeline.
- Short Cycles and Incremental Change. Avoid “big bang” releases and bloated processes. Iterate in short cycles so you can adapt quickly to new and changing needs.
- Always be ready to adapt or implement a new process, to achieve this keep your data analytic pipeline iteration in short cycles which will be easily adapted to incremental changes.
- Enable new methods like object versioning, continuous integration and continuous deployment, automated testing as per the model, and data artifacts requirement.
- Ensure the quality and testing of the model is top on the list and provide no chance for untested models to come for production.



Benefits of DataOps

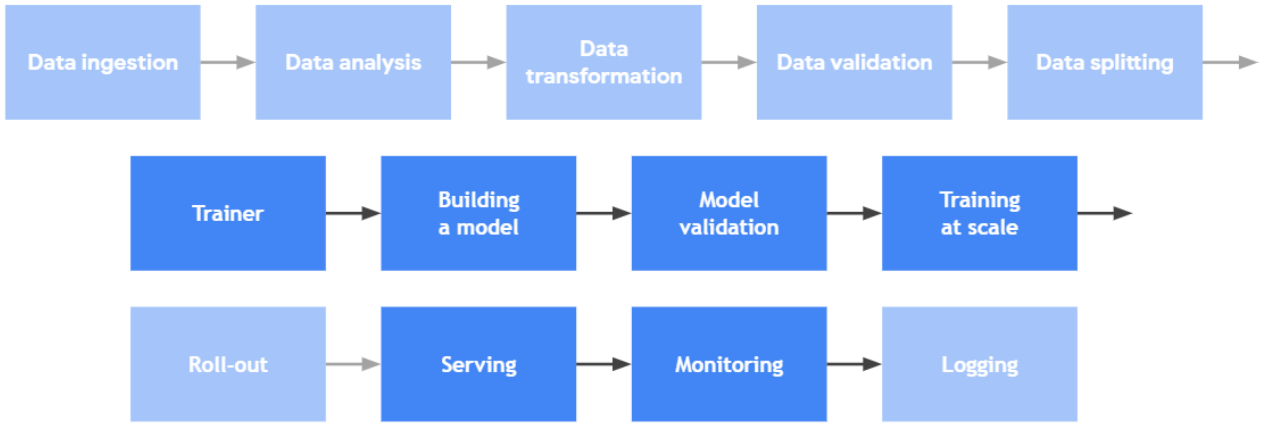
- From DataOps implementation we can overcome the barriers between different pods or departments and functions, through which we will be able to have a good knowledge transfer, reduce delays due to miscommunication between pods, which eventually builds trust, responsibility, collaboration, communication and increases productivity.

- Using DataOps we will be able to streamline and automate the analytic pipelines, through this novel features and insights are available quickly and it reduces manual effort too.

- Implementing DataOps streamlining of the analytic pipelines is easy, by this we achieve speed and robustness of the models. DataOps also monitors the pipelines and helps in identifying the bottlenecks and potential issues through which we will be able to avoid errors in the pipeline.

DATA AND ANALYTICS PIPELINES

Leveraging Kubelow as shown below allows you to operationalize DataOps at scale.



- **Data ingestion**: The first step is to ingest the data into the pipeline, ingested data in this step is used for Data analysis in the next step.

- **Data analysis:** In the Data analysis step we analyze the ingested data to check for anomalies and decide. what kind of transformations to be done for it.

- **Data transformation:** As per the analysis and transformations decided in the Data analysis steps, here we transform the data as per requirement, like applying feature engineering, cleansing of data.

- **Data validation:** After the transformation of data in the Data transformation step, here we validate the transformed data to check whether the applied transformations are proper or any anomalies are present in the data, and to check the data quality. Data splitting: Once we validate the transformed data and the data quality is good then we split the data into 2 or 3 parts as per requirement (train data, test data, and validation data). Model training and tuning:

- **Data ingestion**: The first step is to ingest the data into the pipeline, ingested data in this step is used for Data analysis in the next step.

- **Data analysis:** In this step we analyze the ingested data to check for anomalies and decide. what kind of transformations to be done for it.

- **Data transformation:** As per the analysis and transformations decided in the Data analysis steps, here we transform the data following the requirements, such as applying feature engineering, and cleansing of data.

- **Data validation:** here we validate the transformed data to check whether the applied transformations are proper or any anomalies are present in the data, and to check the data quality.

- **Data splitting:** Once we validate the transformed data and the data quality is good then we split the data into 2 or 3 parts as per requirement (train data, test data, and validation data).

- **Model training and tuning:** As we have split the data into the train and test, we send the data to the model for training. For the better performance and higher accuracy of the model, we do hyperparameter tuning using few libraries like Keras, Katib, etc. The output of this step is a saved model that is used for evaluation, and another saved model that is used for online serving of the model for prediction.

- **Model evaluation and validation:** Once the model training is completed it is exported to evaluate the test data on it to assess the model quality. By this evaluation, we can assure that the model's performance is good or poor and even helps in finding which part of the model is doing great and which is not. It is also used as a benchmark for metrics if it is doing better than the old model or version we can decide to take it for production or not.

- **Model serving for prediction:** Once the new model is validated it will be deployed for microservices for online prediction, it is monitored and logged for further improvements.

The Impact of a Successful DataOps Practice

By DataOps implementation, we can see the improvements all over the data pipelines, so that applying the data change across all the pipelines can happen in minutes whereas before implementation it will take up to weeks or months. Through this, we get optimized reports.

CONCLUSION

DataOps mainly focuses on increasing data consumption in an effective and agile way, by following all other security and governance policies. Companies that have deployed the DataOps successfully have a good track of the data assets available for their access, through which quality of the data can be trusted and make its use to its maximum potential. DataOps is not any specific method or tool, it is a collective principle, for implementing it in an organized way.

References:

1. 2019 Global data management research: Taking control in the digital age." Experian, 2019 link
2. The DataOps CookBook link