indium
Make Technology Work

DataOps in Data Analytics

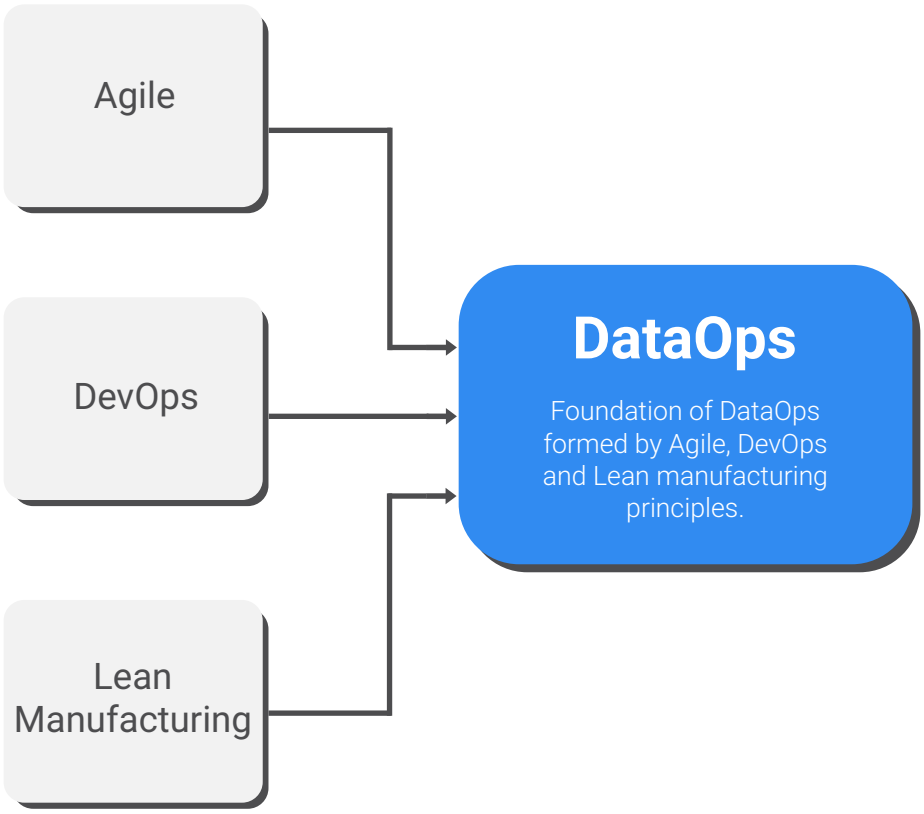
A Whitepaper

Understanding DataOps

Executive Summary

DataOps(Data Operations) is a collaborative data management practices, an integrated, automated and process-oriented used by data and analytics teams.Objective of DataOps is to deliver the high value and manage risks, merge DevOps and agile methodologies to manage data in arrangement with business goals.

The DataOps Methodology is designed 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.



DataOps Architecture Design

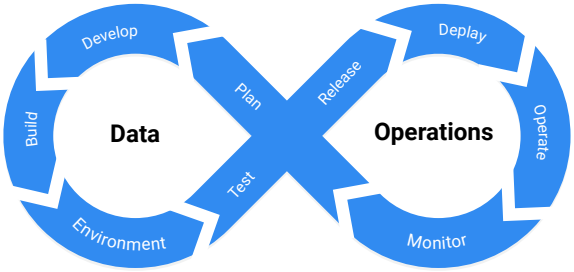
DataOps uses agile practices to orchestrate tools, code, and infrastructure to deliver the trusted data quickly with improved security. This process hepls business to delivery cost effective analytical insights.

DataOps architecture and processes brings new business insights by allowing the rapid development and deployment of innovative, high quality data analytic pipelines.

DataOps Process

Business Needs

- High Quality Data
- Efficient Team Work
- Reliability
- Speed
- Security



Business Values

- Rapid Error Catching
- Realtime Insights
- Efficiency
- Boosted Agility

Starting Point to Assess the DataOps Process

Data Governance and People	• Data Governance • Defined Clear Roles • Use Case for Data Movement • Data Tools • Security and Compliance
Testing & Release	• DTAP Environments • Testing • Build and Deploy Process
Development	• Pipeline and Design Patterns • Centralized Ingestion • Centralized Computation • Data Abstraction • Source Control
Monitor	• Alerting and Remediation • Efficiency • Statistical Process Control

DataOps On Azure Cloud

DataOps methodologies is powerful tool which helps organizations to grow and achieve the cost benefits and high-quality product. Following best practices of DataOps helps the data engineers and business users to understand the data better and relay on the data in trusted manner.

Azure cloud services provide the wide varieties of excellent tools to support the implementation based on it's could service.

DataOps follows certain fundamentals for performance in the cloud environment.

- Automated Governance.
- Multi-dimensional agility.
- Elasticity.
- Flexibility.
- Improved access and integration
- Multi-model data access.
- Increased Coloration
- Insights
- Seamless self-service capabilities for business users

Tools for DataOps Processes (Azure Environment)

Data Processing	Data Storage	Data Governance	Reporting
Apache NiFi			
Azure Data Factory			
Azure Databricks			
Azure Synapse Analytics	Azure Data Lake	Microsoft Purview	Power BI

DataOps Benefits for Business

Every organization has a huge volume of data and not all the data would give the meaningful insight to it's organization. So service providers required to understand the client's needs and build the solution as per the requirements.

YOY data growth would be high and identifying various type of new data sources and storing them is very critical. The main issues is availability of data and make them available to number of users.

Cloud services will handle this type of situation effectively and enables data storage for high data growth

Some of the main advantages of using data platform on clouds

- No more silos of data.
- Central Repository for source data, i.e., Data Lakes on Clouds.
- Secured data storage in standard format.
- Being able to Grow to any scale with as low costs as possible.
- To predict future outcomes.

DataOps for the Modern Data Warehouse

A modern data warehouse (MDW) lets you easily bring all of your data together at any scale. It doesn't matter if it's structured, unstructured, or semi-structured data.

You can gain insights to an MDW through analytical dashboards, operational reports, or advanced analytics for all your users.

Setting up an MDW environment for both development (dev) and production (prod) environments is complex.

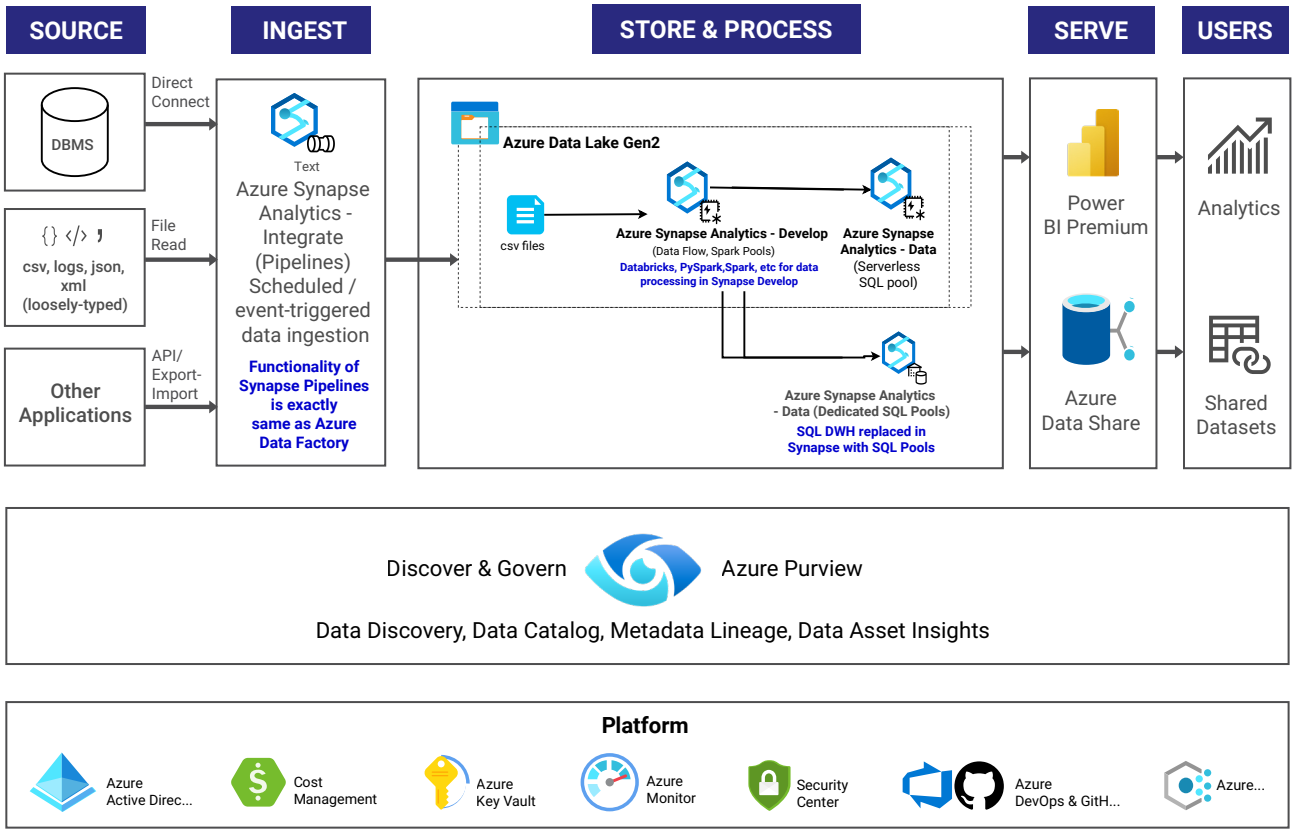
Automating the process is key. It helps increase productivity while minimizing the risk of errors.

Typical Soultion Requirement

Data Collection from different sources

- Automated deployment
- Implement (CI/CD) pipeline
- Source control for code management
- Carry out integration tests
- Run pipelines on a scheduled basis.
- Support future agile developmen
- Data and database Security
- Support concurrent users
- Azure Key Vault for configuration

Reference Architecture



Dataflow

Azure Synapse pipeline orchestrates and Azure Data Lake Storage (ADLS) Gen2 stores the data

- Pipeline copy job transfer the data into Gen2
- Synapse Analytics Develop (dataflow, notebook (PySpark)) cleanse and standardize the data.
- Synapse Analytics Develop transform step that converts the data into a format that you can store in the data warehouse.

- The pipeline serves the data in two different ways

- Data Share delivers the snapshot of data for downstream process and system
- Polybase moves the data from the data lake to Power BI accesses the data and presents it to the business use

Deployment

The following list contains the high-level steps required to set up the solution with corresponding Build and Release Pipelines:

- **Initial setup:** Install any prerequisites, import the Azure Samples GitHub repository into your own repository, and set required environment variables.

- **Deploy Azure resources:** The solution comes with an automated deployment script. It deploys all necessary Azure resources and Microsoft Azure Active Directory service principals per environment. The script also deploys Azure pipelines, variable groups, and service connections.

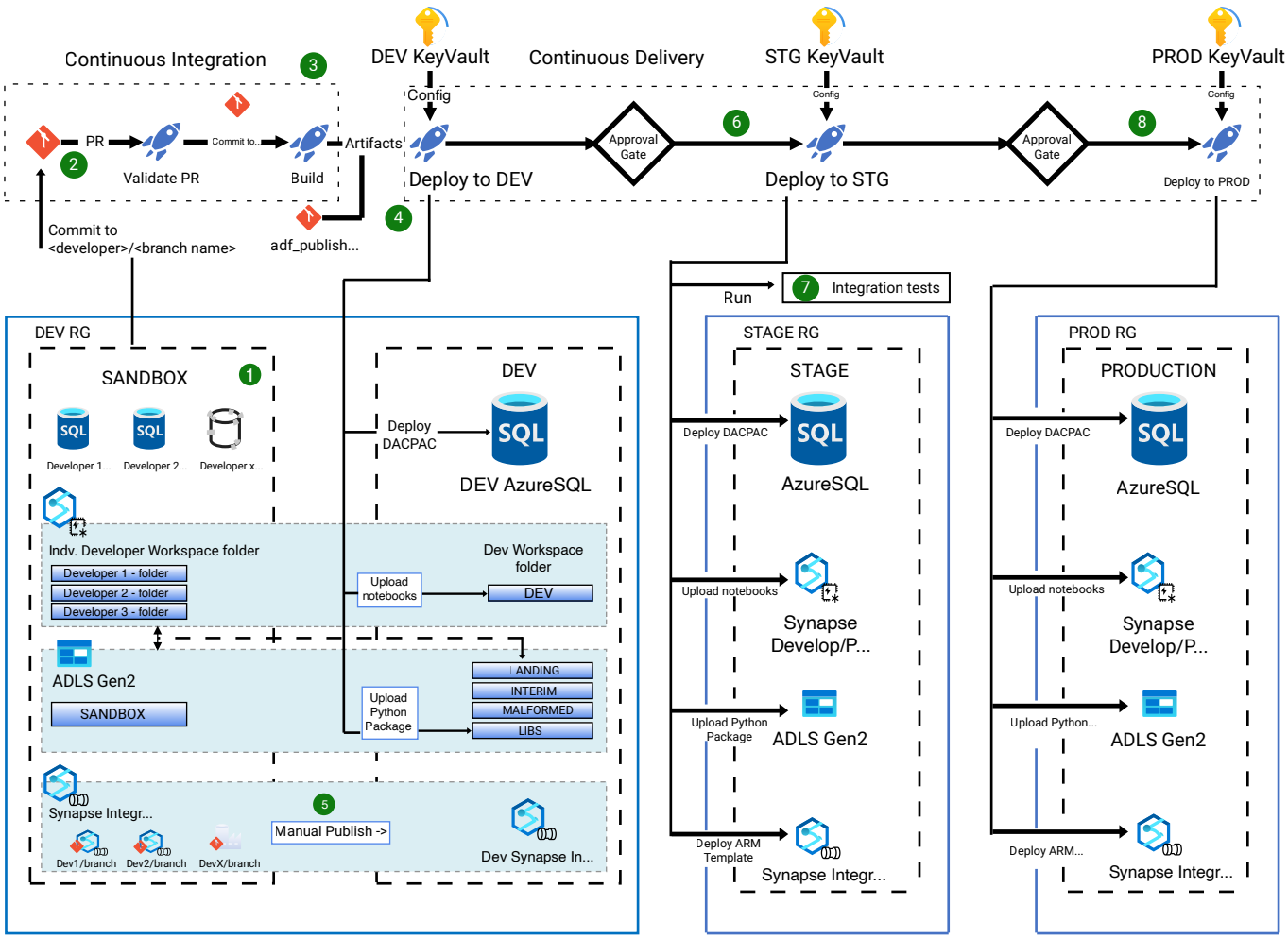
- **Set up git integration in dev Synapse Pipleline:** Configure git integration to work with the imported GitHub repository.

- **Carry out an initial build and release:** Create a sample change in pipeline, like enabling a schedule trigger, then watch the change automatically deploy across environments.

If deployment is successful, there should be three resources groups in Azure representing three environments:

development, staging and production. There should also be end-to- end build and release pipelines in Azure DevOps that can automatically deploy changes across these three environments.

CI/CD Archtecture



- Developers commit the changes into branch.
- Pull request(PR) automatically kicks-off of the validation,testing and builds then publishes the build artifacts into dev except synapse pipeline.
- Developers manually publish to the dev synapse pipeline from the collaboration branch (main).

  The manual publishing updates the Azure Resource Manager (ARM) templates in the adf_publish branch.

- On Approval, the release pipeline continues with the second stage, deploying changes to the stage environment.
- Run integration tests to test changes in the stage environment.
- Upon successful completion of the second stage, the pipeline triggers a second manual approval gate.
- On Approval, the release pipeline continues with the third stage, deploying changes to the production environment.

The Final Word

According to Experian global data management report the following are the key findings

- All Business are looking to leverage data to better understand their customer, but the definition of the customer varies across the business.
- Process around date need to be structured enough to establish trust, but flexible enough to modify the view based on the context of the data usage.
- Trusted data enables a host of business benefits, but management practices have not kept up with changing data usage.
- Organisations need to invest in data management to drive innovation and keep up with data explosion.

DataOps solution would address most of the data challenges for any organization

References

https://learn.microsoft.com/ https://www.experian.co.uk/

USA
Cupertino | Princeton
Toll-free: +1-888-207-5969

INDIA
Chennai | Bengaluru | Mumbai | Hyderabad
Toll-free: 1800-123-1191

UK
London
Ph: +44 1420 300014

SINGAPORE
Singapore
Ph: +65 6812 7888

www.indiumsoftware.com

For Sales Inquiries
sales@indiumsoftware.com

For General Inquiries
info@indiumsoftware.com