Discovering DataOps: A Comprehensive Review of Definitions, Use Cases, and Tools

Authors:
Kiran Mainali
Lisa Ehrlinger
Johannes Himmelbauer
Mihhail Matskin

Presenter:
Kiran Mainali
KTH Royal Institute of Technology
mainali@kth.se

Presenter Bio

Contact: mainali@kth.se

Kiran Mainali received his Masters Degree in ICT Innovation from KTH Royal Institute of Technology and Technische Universität Berlin, specializing in Cloud Computing and Services in February 2014. He also holds a Bachelor of Business Information System (2014) from Kathmandu University, Nepal.

Kiran has a tech entrepreneurial background with eight years (2010-2018) of involvement in Spark Technology Pvt. Ltd. as Co-founder. He worked as a technology transformation consultant for various organizations and projects. He led and established his startup to an established Digital Agency in Nepal as COO and Software Engineer.

His current career focuses on exploring new trends and technologies in Cloud Computing and Data Science to ensure the benefits of such tools and technologies in practical projects. He is actively looking for new career opportunities as a researcher either in academia or in the industry to further contribute to the field.

Introduction

- Data collection and analysis approach has changed.
- Data analytics process is becoming complex.
- Data management is vital due to rapid generation and availability of various format.
- DevOps is proven to be successful in SDLC but in Data Analytics, challenges cannot be simply solved by exploiting DevOps.
- DataOps is an emerging approach to execute data analytics projects.
- DataOps Manifesto provides principles that summarizes the best practices, goals, philosophies, mission and values.
- DataOps promises to streamline the process of the build, change, and manage data pipelines.
- Despite few scientific contributions, numerous commercial resources are available.

Research Aims

- Uncover DataOps from the scientific perspective with rigorous review of research and tools to:
  - Outline definitions of DataOps and their ambiguities and challenges of implementation.
  - Identify the extent to which DataOps covers different stages of data lifecycle.
  - Provide a comprehensive overview on tools and their suitability for different stages in DataOps.

Research Method

► **Explorative Qualitative Process:** Comprehensive literature study of the data lifecycle, DataOps, DevOps, Agile, Lean manufacturing, data governance, data lineage and provenance, data pipeline and featured based comparison of tools and technologies.

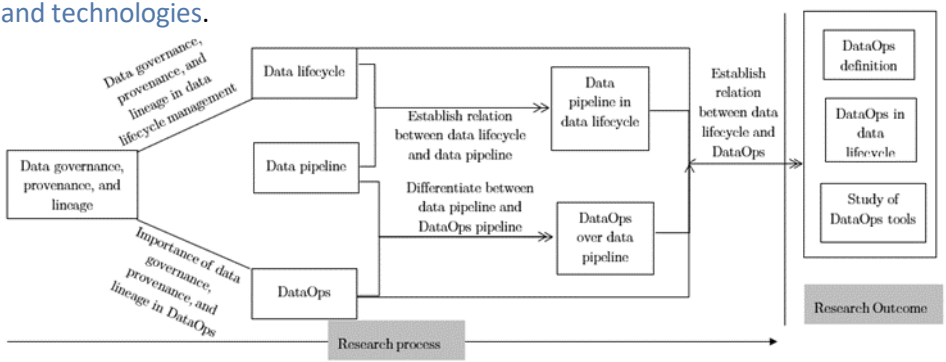

Figure: Illustration of the exploratory research method

Result and Evaluation

DataOps Definition - (1/2)

"DataOps can be defined as data pipeline development and execution methodology by assembling people and technology to deliver better results in a shorter time. With DataOps, people, processes, and technology are orchestrated with a degree of automation to streamline data flow from one stage of the data lifecycle to another. DataOps using Agile, DevOps, and SPC's best practices in combination with technologies, and processes promotes data governance, continuous testing and monitoring, optimization on the analysis process, communication, collaboration, and continuous improvement."

DataOps Definition - (2/2)

- DataOps has its own approaches on top of derived processes from other methodologies.
  - Separating the production environment from development gives room for data workers to experiment with changes and reduce fear of failure.
  - Product quality can be assured by continuous testing and cross-environment monitoring.
  - Including customers and other stakeholders in data analytics project sets communication and feedback process to minimum iteration.

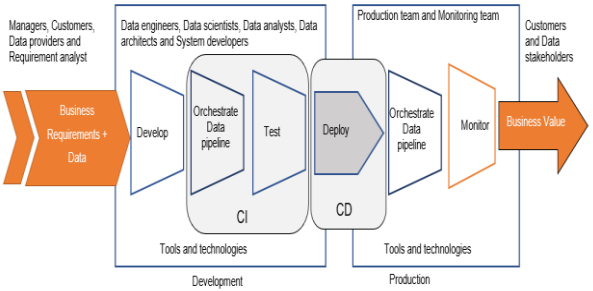

Figure: DataOps pipeline

DataOps in Data Lifecycle

- Data pipeline transport data from one stage of lifecycle to another.
- DataOps restructures data pipelines and take them out of the black box making them measurable, maintainable through collaboration, communication, integration, and automation.
- Goal is to minimize analytics cycle time.
- Applies to entire data lifecycle.

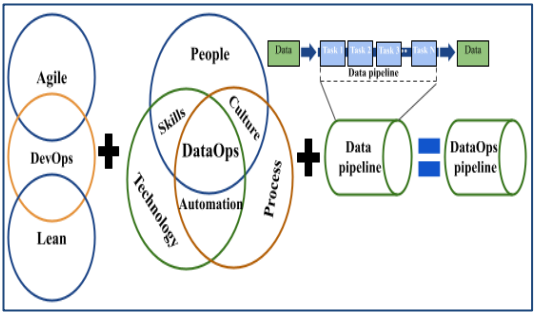

Figure: Data Pipeline to DataOps

- Collaborates people and tools to better manage data lifecycle.
- Quality assurance and the DataOps' principle of reproducible and reuse are highly dependent on managing and maintaining data lifecycle change events.
- DataOps utilizes the technical modularity of orchestration, workflow management, and automation tools.

Ambiguities in DataOps

- DataOps is just DevOps applied in data analytics.
- DataOps is all about using tools and technology in data pipeline.
- DataOps is an expensive methodology to implement.
- With DataOps, there is no need of coding.
- DataOps can only be used on data analysis tasks.
- DataOps and data pipeline are two different ways of data analytics project implementation.

Challenges in DataOps Implementation

- Changing the organization's culture.
- Innovation with low risk.
- Cost of DataOps.
- Transition from expertise-based team to cross-functional teams.
- Managing multiple environments.
- Sharing Knowledge.
- Tools and technology diversity.
- Security and quality.

Evaluation of DataOps Tools and Technologies – (1/8)

- Tools are categorized into various functional area.
- Selection of tools is based on mass user base, relevant features to support functional area and popularity.
- A baseline to further research for selecting tools in DataOps tasks.
- Tools presented are compared based evaluation criteria.

Evaluation Criteria	
Complexity	▪ **HIGH:** Need a high level of coding and configuration to install the product. ▪ **MEDIUM:** Moderate level of coding and configuration required. ▪ **LOW:** Easy to install with no line of code or a few lines of code.
Usability	▪ **HIGH:** Easy to use with little or no technical, coding, or system-related knowledge. ▪ **MEDIUM:** Moderate knowledge of the system, code architecture, or technical detail is required. ▪ **LOW:** High level of technical expertise and/or coding knowledge is required.
Compatibility	▪ **HIGH:** Supports a wide range of tools and operation environment ▪ **MEDIUM:** Have some level of support. ▪ **LOW:** Little or no support available.
Application	▪ **GENERIC:** Can be used in a variety of projects based on the nature of tools. ▪ **SPECIFIC:** Industry/project-specific usage.
Lifecycle	**Lists in which data lifecycle stage the tool can mostly be used.**
License	**Describes whether the tool is commercial, opensource, freemium, free + commercial and other pricing forms.**

Evaluation of DataOps Tools and Technologies – (2/8)

1. Workflow Orchestration tools

Tools	Lifecycle	Complexity	Usability	Compatibility	Application	License
Airflow	Creation/collection, Process, Analyze	HIGH	MEDIUM	HIGH	GENERIC	Opensource
Apache Oozie	Creation/collection, Process, Analyze	HIGH	MEDIUM	LOW	GENERIC	Opensource
Reflow	Process, Analyze	HIGH	LOW	LOW	SPECIFIC	Opensource
DataKitchen	Process, Analyze	LOW	HIGH	HIGH	GENERIC	Commercial
BMC Control-M	Process, Analyze	MEDIUM	MEDIUM	HIGH	GENERIC	Commercial
Argo Workflows	Process, Analyze	HIGH	LOW	LOW	GENERIC	Opensource
Apache NIFI	Creation/collection, Process, Analyze	MEDIUM	MEDIUM	MEDIUM	SPECIFIC	Opensource

Evaluation of DataOps Tools and Technologies – (3/8)

2. Testing and monitoring tools

Tools	Lifecycle	Complexity	Usability	Compatibility	Application	License
iCEDQ	Creation/collection, Storage, Analyze	LOW	HIGH	HIGH	GENERIC	Commercial
Data Band	Process	HIGH	LOW	MEDIUM	GENERIC	Opensource + Commercial
RightData	Storage, Analyze, Process	MEDIUM	MEDIUM	HIGH	GENERIC	Commercial
Naveego	Creation/collection, Process, Storage	HIGH	HIGH	LOW	SPECIFIC	Commercial
DataKitchen	Creation/collection, Process, Storage	HIGH	MEDIUM	HIGH	GENERIC	Commercial
Entreprise Data Foundation	Storage, Analyze Process	HIGH	LOW	LOW	SPECIFIC	Free Non-profit

Evaluation of DataOps Tools and Technologies – (4/8)

3. Deployment automation tools

Tools	Lifecycle	Complexity	Usability	Compatibility	Application	License
Jenkins	Collection/creation, Process, Analyze Publish, Storage	MEDIUM	HIGH	HIGH	GENERIC	Opensource
DataKitchen	Collection/creation, Process, Analyze Publish, Storage	HIGH	MEDIUM	HIGH	GENERIC	Commercial
Circle CI	Collection/creation, Process, Analyze Publish, Storage	MEDIUM	MEDIUM	MEDIUM	GENERIC	Free + Commercial
GitLab	Collection/creation, Process, Analyze Publish, Storage	MEDIUM	MEDIUM	HIGH	GENERIC	Opensource + Commercial
Travis CI	Collection/creation, Process, Analyze Publish, Storage	MEDIUM	HIGH	HIGH	GENERIC	Free + Commercial
Atlassian Bamboo	Collection/creation, Process, Analyze Publish, Storage	LOW	HIGH	HIGH	GENERIC	Commercial

Evaluation of DataOps Tools and Technologies – (5/8)

4. Data governance tools

Tools	Lifecycle	Complexity	Usability	Compatibility	Application	License
Apache Atlas	Collection/creation, Process, Analyze Publish, Storage	HIGH	MEDIUM	MEDIUM	GENERIC	Opensource
Talend	Collection/creation, Process, Analyze Publish, Storage	MEDIUM	MEDIUM	MEDIUM	SPECIFIC	Opensource + Commercial
Collibra	Collection/creation, Process, Analyze Publish, Storage	LOW	LOW	LOW	SPECIFIC	Commercial
IBM	Collection/creation, Process, Analyze Publish, Storage	MEDIUM	HIGH	MEDIUM	GENERIC	Commercial
OvalEdge	Collection/creation, Process, Analyze Publish, Storage	LOW	HIGH	HIGH	GENERIC	Commercial

Evaluation of DataOps Tools and Technologies – (6/8)

5. Code, artifact and data versioning tools

Tools	Lifecycle	Purpose	License
GitLab	Collection/creation, Process, Analyze, Publish, Storage	Code versioning	Free + Commercial
GitHub	Collection/creation, Process, Analyze, Publish, Storage	Code versioning	Free + Commercial
DVC	Collection/creation, Process, Analyze, Publish, Storage	Data versioning	Opensource
Docker Hub	Collection/creation, Process, Analyze, Publish, Storage	Docker Image versioning	Free + Commercial

- Additional Information:
  - Tools are used in **GENERIC** applications
  - Complexity and Usability are **MEDIUM**
  - Compatibility **is HIGH**

Evaluation of DataOps Tools and Technologies – (7/8)

6. Collaboration and communication tools

Tools	Lifecycle	License
Slack	Collection/creation, Process, Analyze, Publish, Storage	Freemium + User-based pricing
Jira	Collection/creation, Process, Analyze, Publish, Storage	Freemium + User-based pricing
Trello	Collection/creation, Process, Analyze, Publish, Storage	Freemium + User-based pricing

- Additional Information:
  - No installation required
  - Web-based with elegant user interface
  - Widely used in any organizations

7. Analytics and visualization tools

Tools	Lifecycle	Complexity	Usability	Compatibility	License
Tableau	Analyze	LOW	MEDIUM	HIGH	Commercial
Power BI	Analyze	LOW	MEDIUM	MEDIUM	Commercial
QlikView	Analyze	LOW	MEDIUM	LOW	Commercial

- Additional Information:
  - Tools are used in **GENERIC** applications

Evaluation of DataOps Tools and Technologies – (8/8)

► DataOps takes advantage of existing tools and technology.

► Hundreds of tools available in the market with similar features and functionalities.

- Choosing right from the bucket needs informed decision.

► All stages of data lifecycle are well covered by combinations of tools and technologies.

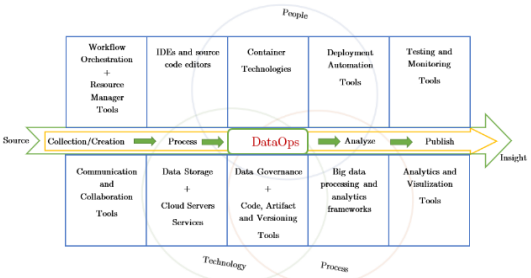

Figure: DataOps framework

Conclusions

- Data analysis itself is a broad field, where numerous tools, approaches, and technologies can lead to the same result.
- DataOps advocates collaboration, quality control, and fast delivery of analysis tasks by extending proven DevOps methodology from SDLC as well as Agile and Lean Manufacturing's SPC.
- DataOps is data pipeline execution methodology by assembling people and technology to deliver better results in a shorter time.
- With DataOps, people, processes, and technology are orchestrated with a degree of automation to streamline data flow from one stage of the data lifecycle to another.
- Selection of right tool from pool of available tools for right task is work of proper research and planning.
- Using suitable tools allows to coverall stages of the data lifecycle with the DataOps methodology.

Future Work

- Experiment on DataOps approach in different data analytics projects to:
  1. Validate the efficacy of the methodology itself
  2. Measure the performance of different tools in various use cases.

- A compatibility rating (based on combined performance when used together in data analytics task) of one tool from one functional group to other functional groups would help DataOps practitioners make informed decisions.

Acknowledgement

The research presented here has been funded by the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Federal Ministry for Digital and Economic Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET -- Competence Centers for Excellent Technologies Programme managed by Austrian Research Promotion Agency FFG. And is further supported by the EC H2020 project "DataCloud: Enabling the Big Data Pipeline Lifecycle on the Computing Continuum"(Grant nr. 101016835).

scch { software competence center hagenberg }

DataCloud

Thank you