Morris & Opazo

aws partner network

Advanced
Consulting
Partner

Data & Analytics
Competency

Channel Partner

EBOOK:

Intro to DataOps

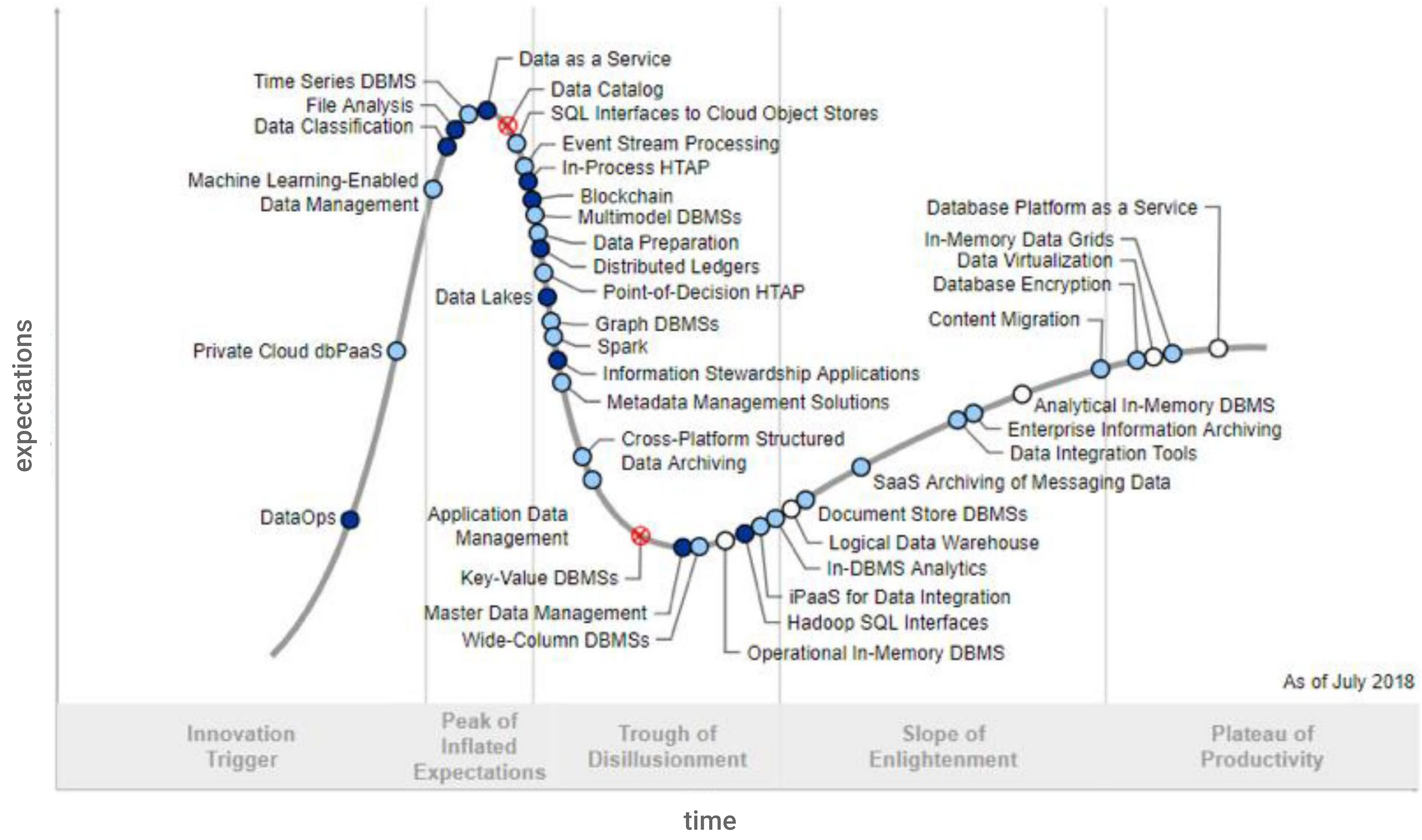
We Love the Cloud

www.morrisopazo.com contacto@morrisopazo.com

Contents

...DATA is now not only considered as an Asset for COMPETITIVE ADVANTAGE; but now a Strategic Asset for COMPETITIVE SURVIVAL...

Gartner Hype Cycle for Data Management Technologies, 2018



Plateau will be reached:

○ less than 2 years ● 2 to 5 years ● 5 to 10 years △ more than 10 years ⊗ obsolete before plateau

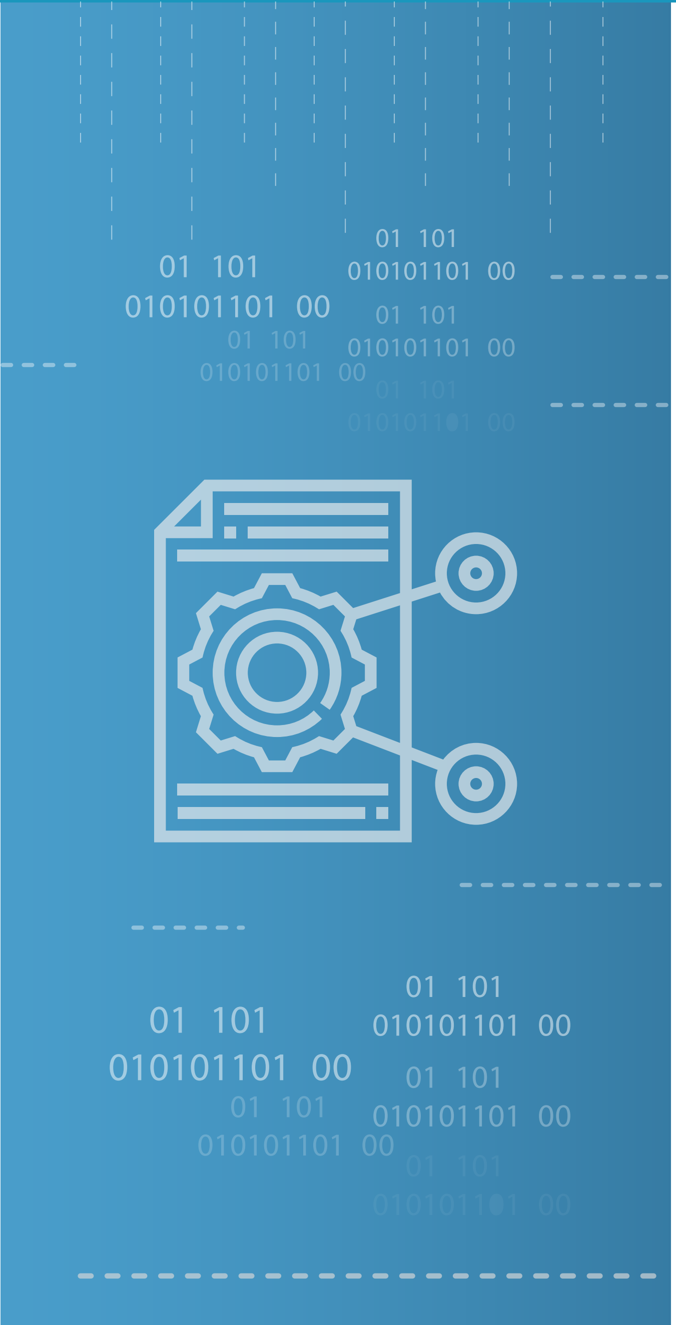
Executive Summary

DataOps is an emerging set of Practices, Processes, and Technologies for building and enhancing data and analytics pipelines to meet business needs quickly. As these pipelines become more complex and development teams grow in size, organizations need better collaboration, development and operations processes to govern the flow of data and code from one step of the data lifecycle to the next – from data ingestion and transformation to analysis and reporting. **The goal is to increase agility and cycle times**, while reducing data defects, increasing application reliability and giving developers and business users greater confidence in data analytics output.

DataOps builds on concepts popular in the software engineering field, such as **agile, lean, and continuous integration/continuous delivery**, but addresses the unique needs of data and analytics environments, including the use of multiple data sources and varied use cases that range from data warehousing to data science. It relies heavily on **test automation, code repositories, collaboration tools, orchestration, monitoring frameworks, and workflow automation to accelerate delivery times while minimizing defects.**

DataOps requires cultural shift. It is not something that can be implemented all at once or in a short period of time. DataOps is a journey. Leaders use productivity metrics to gauge their progress and impel them and their teams to continuously search for new ways to cut wasted effort, streamline steps, automate processes, increase output, and get it right the first time. For large organizations with big development teams, DataOps is an antidote to many of the woes that beset IT and development organizations.

Why do we need DataOps?



Every company is now a software company. The digital economy has created an **unquenchable thirst for data** across all aspects of business.

Data, and access for those that need it, is a **competitive advantage**. Those that can leverage data to drive innovation will win; those that can't, will lose.

Data friction is caused when **constraints** on data prevent people from meeting the ever-growing **demands of the business.**

When data friction becomes the **blocker to innovation**, customers leave, competitors win, and businesses spend more time reacting instead of leading.

What is DataOps?

DataOps is the alignment of **People, Process, and Technology** to enable the rapid, automated, and secure management of data. Its goal is to improve outcomes by bringing together those that need data with those that provide it, eliminating friction throughout the data lifecycle.

It can be summarized as an **Agile Methodology for Data-Driven Organizations.**

Key points:

> **Agile software development** helps deliver new analytics faster and with higher quality.

> **DevOps** automates the deployment of new analytics and data.

> Statistical process controls, used in **lean manufacturing**, test and monitor the quality of data flowing through the data-analytics pipeline.

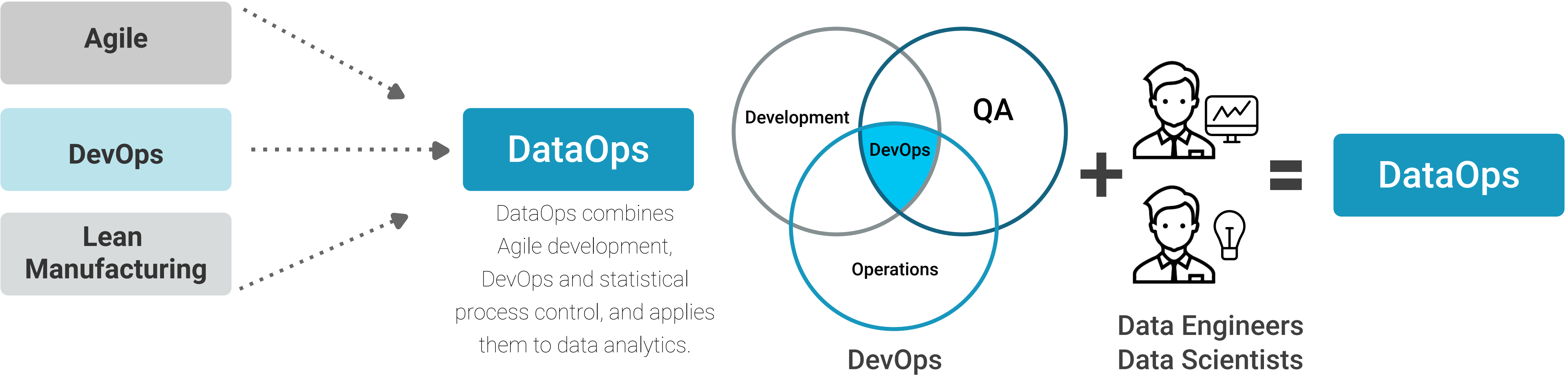
What is DataOps?

DataOps is an integrated approach for delivering **data analytic solutions** that uses **automation, testing, orchestration, collaborative development, containerization, and continuous monitoring** to continuously accelerate output and improve quality.

The purpose of **DataOps is to accelerate the creation of data and analytics pipelines**, automate data workflows, and deliver and operate high-quality data analytic solutions that meet business needs as fast as possible.

“DataOps consists of a stream of steps required to deliver value to the customer. Automate those steps where possible, minimize waste and redundancy, and foster a culture of continuous improvement.”

What is DataOps?



Agile

DevOps

Lean Manufacturing

DataOps

DataOps combines Agile development, DevOps and statistical process control, and applies them to data analytics.

Development

QA

DevOps

Operations

DevOps

+

=

Data Engineers
Data Scientists

DataOps

Key Points of Data Friction

DataOps also mandates a comprehensive technology approach that eliminates key points of friction across:

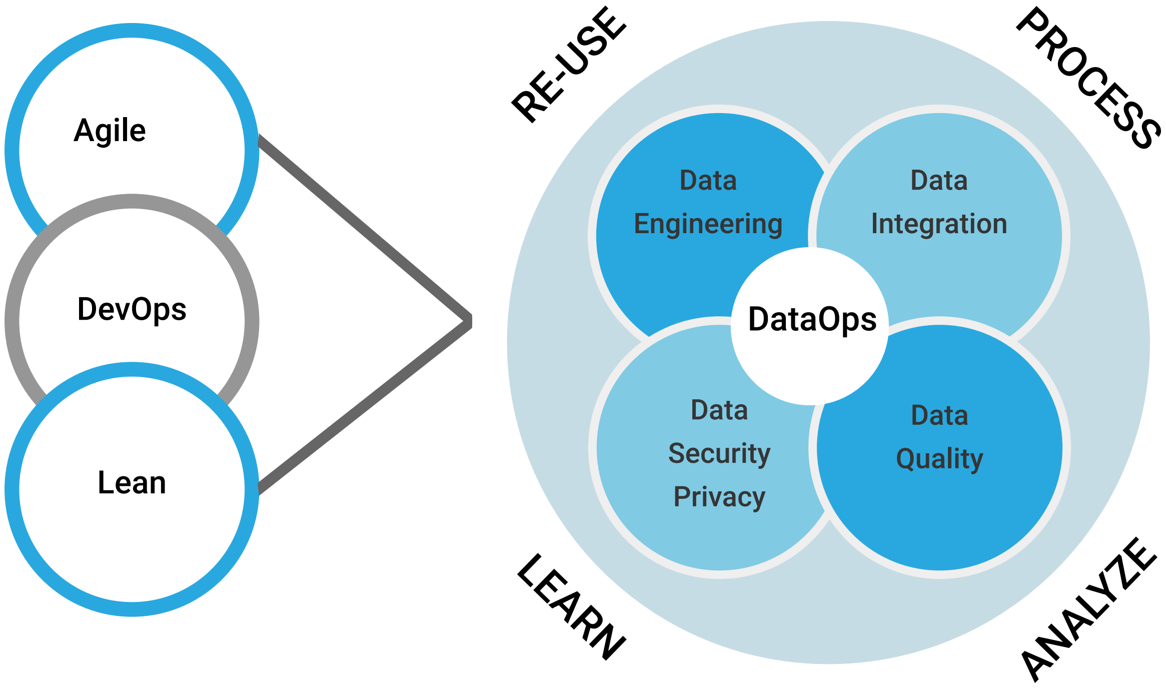
> **Governance :** Security, quality, and integrity of data, including auditing and access controls.

> **Operation:** Scalability, availability, monitoring, recovery, and reliability of data systems.

> **Delivery:** Distribution and provisioning of data environments.

> **Transformation:** Modification of data, including masking and platform migration.

DataOps in the Enterprise



DevOps / DataOps:
Different People and Expectations

DevOps Users & Tools

Software Engineers, comfortable with **coding** and complexity of multiple languages, tools, and hardware/software.

DevOps Users & Tools

Data Scientists, Engineers, and Analysts who want to just analyze **data** and build models - everything else is unwanted complexity.

Introducing the Chief Data Officer...

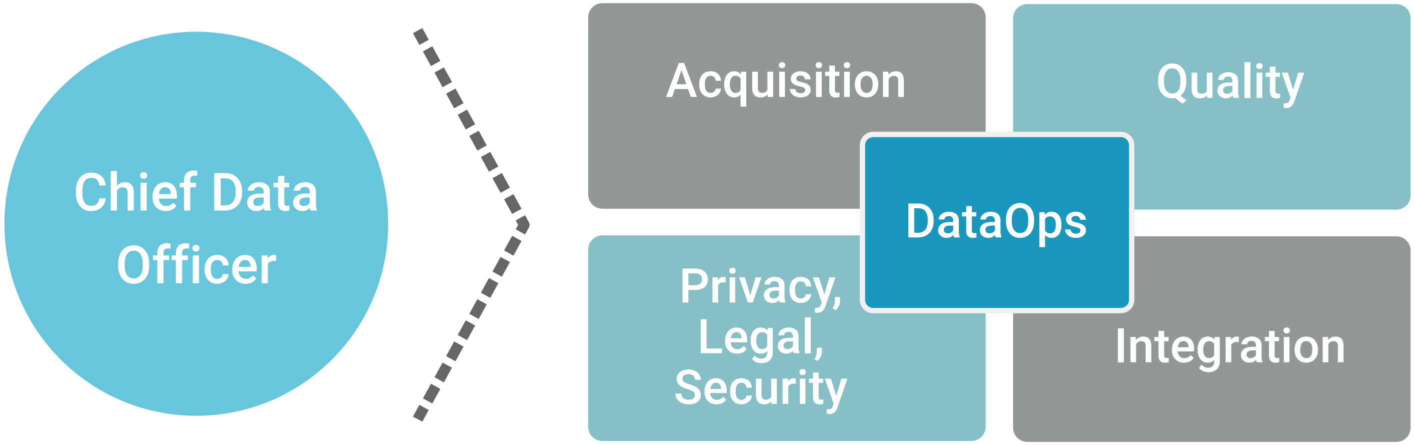
Chief Information Officer

Chief Data Officer

Line of Business Executive

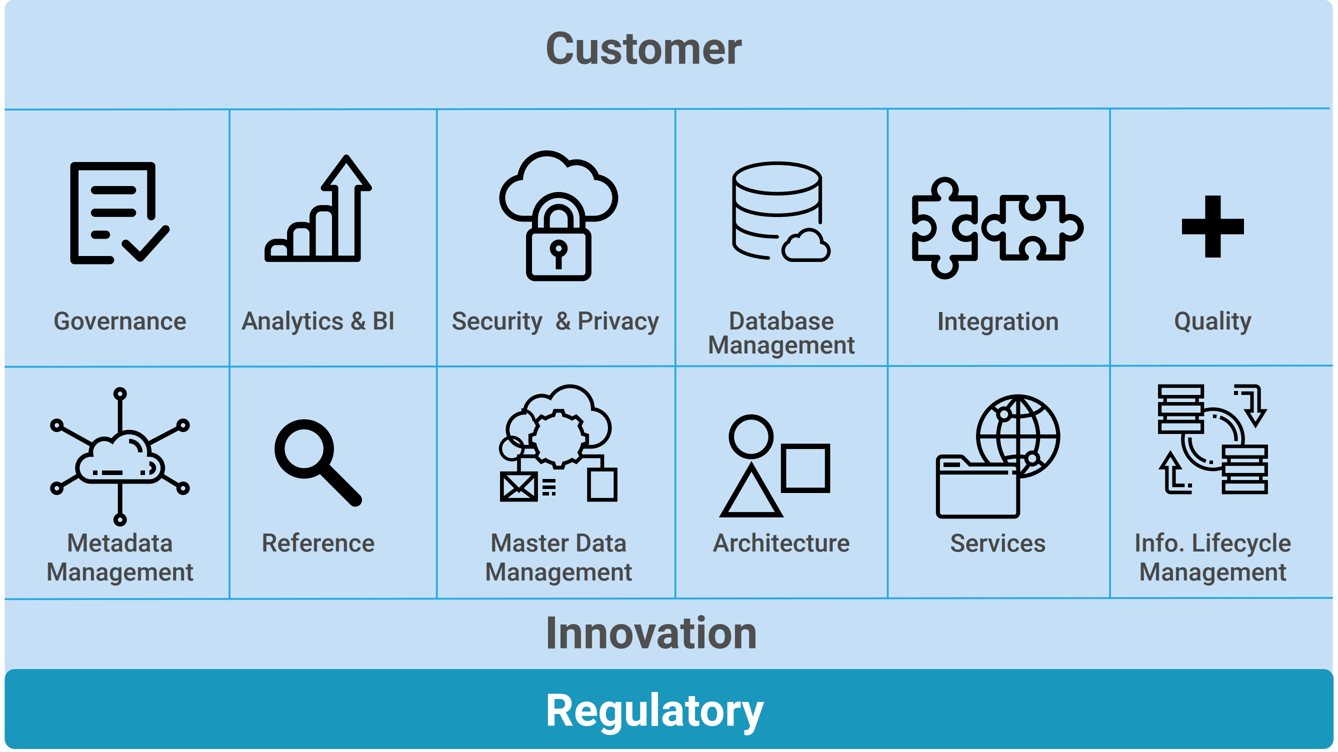
Chief Data Officer

Chief Data Officer will lead the transformation of the **Business Data Environment** to enable DataOps...



Chief Data Officer

Chief Data Officer needs to ensure all core data components are supportable within the DataOps framework...



Customer

Governance

Analytics & BI

Security & Privacy

Database Management

Integration

Quality

Metadata Management

Reference

Master Data Management

Architecture

Services

Info. Lifecycle Management

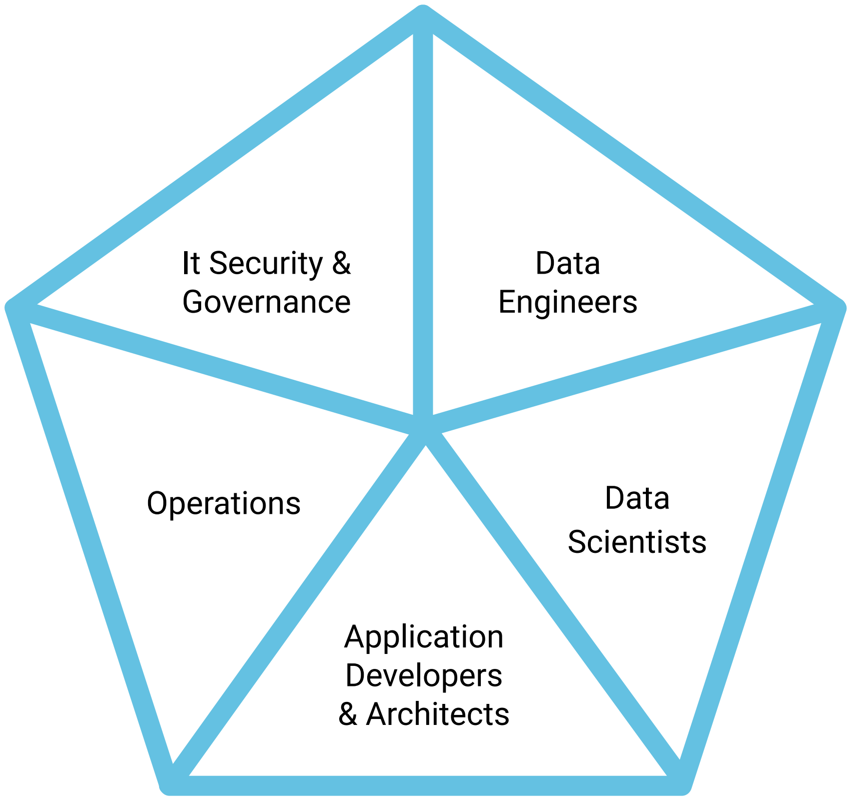
Innovation

Regulatory

Data Team Roles

Roles	Other Job Titles	Responsibilities	Skills
Data Engineer	Database Architect, Data Modeler, Database Administrator, Data QA Engineer, ETL Engineer	Data lakes, Data warehouses, Data marts, Schema design	Databases, Programming, Cloud Infrastructure, Simple storage
Data Analyst	Data Visualization Designer, Business Data Analyst, BI Tableau Developer, Reporting Analyst, Business Intelligence Engineer	Visualizations: Charts, Graphs, Dashboards, Tables, Reports	Programming Statistics, Machine Learning, Data cleaning, Data visualization
Data Scientist	Machine Learning Researcher, Machine Learning Engineer, Quantitative Analyst, AI Programmer Actuary	Algorithms, Models	Domain subject matter expert, Advanced mathematics, Machine learning, Data mining tools, Programming
DataOps Engineer		Orchestrating the analytic pipeline, Promoting features to production, Automating quality	Agile Development, DevOps, Statistical Process Control

Cross-Functional Collaboration



A DataOps methodology requires cross-functional collaboration.

Benefits of DataOps

> **Accelerate Time to Production.** A major driver for DataOps is speed. The idea of streamlined and largely automated analytics pipelines helps deliver new features and insights quickly and reduces manual effort. Moreover, the short feedback and testing cycles help speed up reactions to changing business requirements and increase flexibility.

> **Increase Quality, Reliability and Visibility.** Well-defined analytics pipelines enhance both speed and robustness. For instance, multiple stages of automated and manual tests prevent the deployment of flawed updates. Besides, DataOps also includes monitoring of production environments to identify bottlenecks or potential issues and thereby improves efficiency and stability of infrastructure and applications. Lastly, the convergence of different roles helps align changes throughout various stages, such as when a data engineer is informed about the later cleansing issues encountered by a data scientist or the lack in performance of an ETL process in production. A way to achieve this, can be a Self-Service Application Performance Management Platform that allows all stakeholders to understand and rationalize the performance of analytic applications.

> **Security.** With a unified data platform, organizational data access and privacy policies can be enforced holistically across organizations. Model development and application deployment activities inherit from the data access policies specified by the governance group.

DataOps Goals

The goal of DataOps is to reduce the cost of asking new questions and accelerate the speed of ideas. Using DataOps, organizations can be much more creative because ideas are easier to vet and implement. These organizations will make better decisions, more quickly, increasing their probability of success.

The goals of a DataOps practice are:

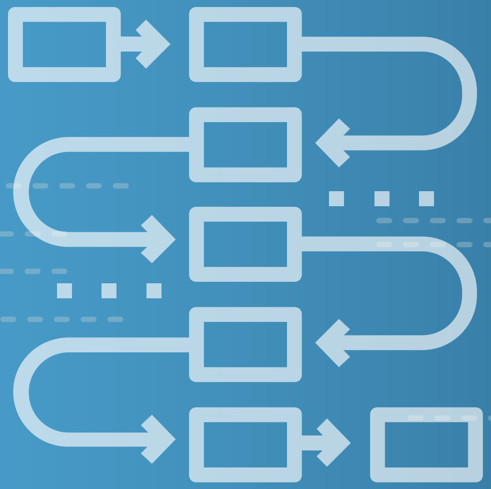
> Continuous model deployment

> Promote repeatability

> Promote agility

> Promote self-service

Best Practices

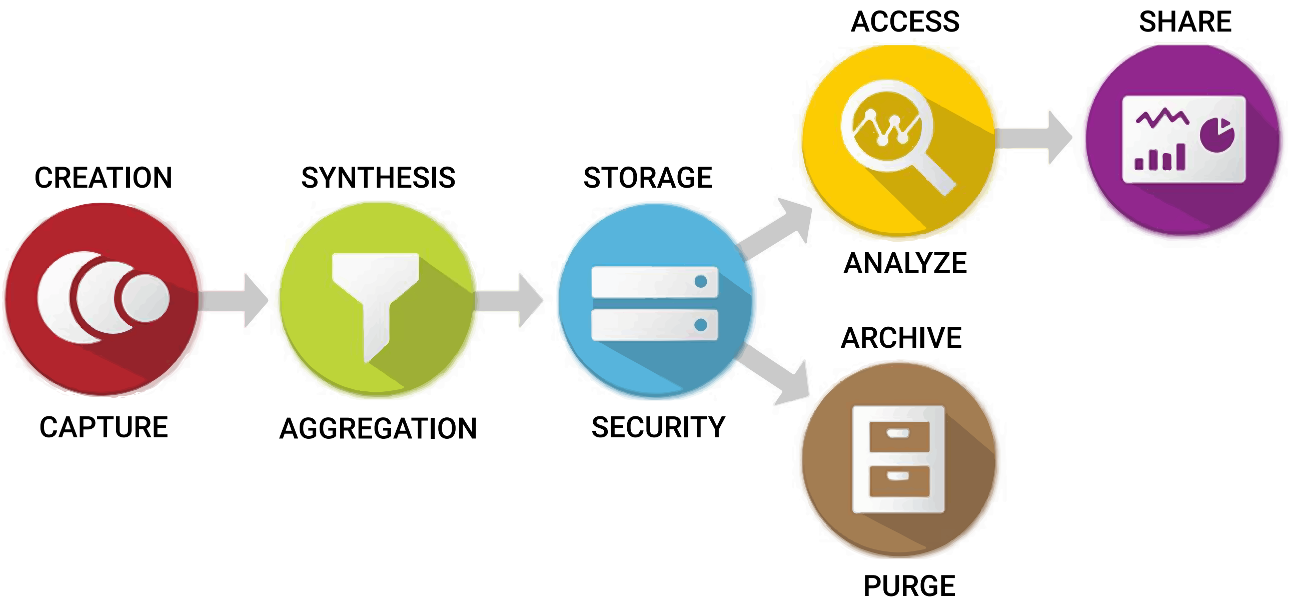


> **Culture.** The core of DataOps is a culture of collaboration and trust. All stakeholders must work together and feel responsible for the entire process. Awareness of the business requirements in all stages is essential.

> **Processes.** DataOps requires well-defined processes, roles, guidelines, and metrics to reinforce DataOps principles. Consequently, many companies establish testing and certification programs to educate and train staffers.

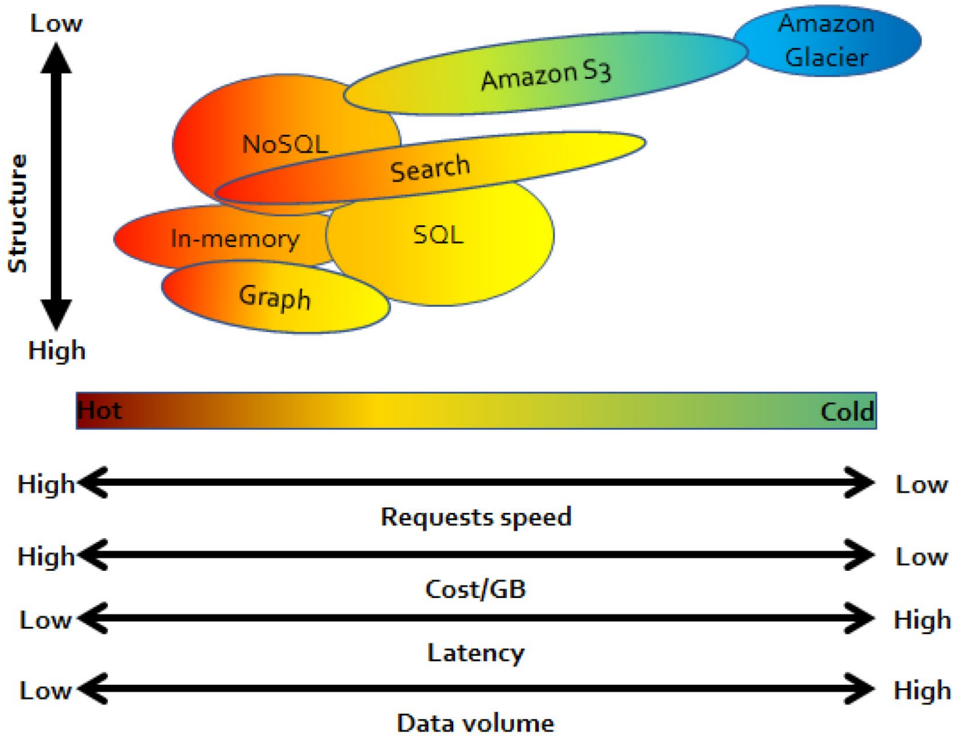
> **Technology.** DataOps requires tools and infrastructure to support automation, testing, monitoring, and orchestration, as well as collaboration and communication among all stakeholders.

Data Lifecycle Management (DLM)

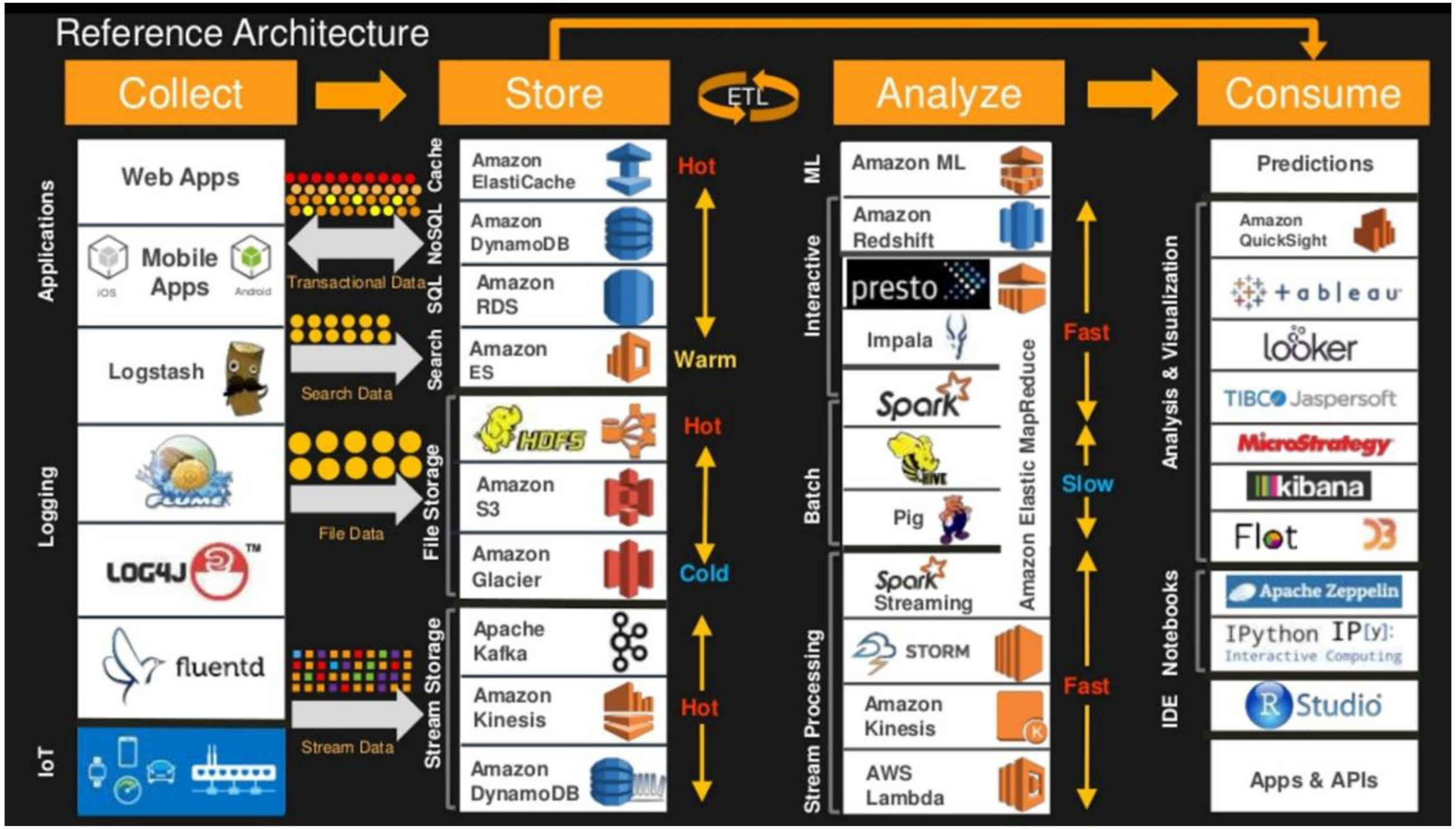


Data Lifecycle Management (DLM) is a process that helps organizations manage the flow of data throughout its lifecycle from creation, to use, to sharing, archive and deletion.

Data Temperatures

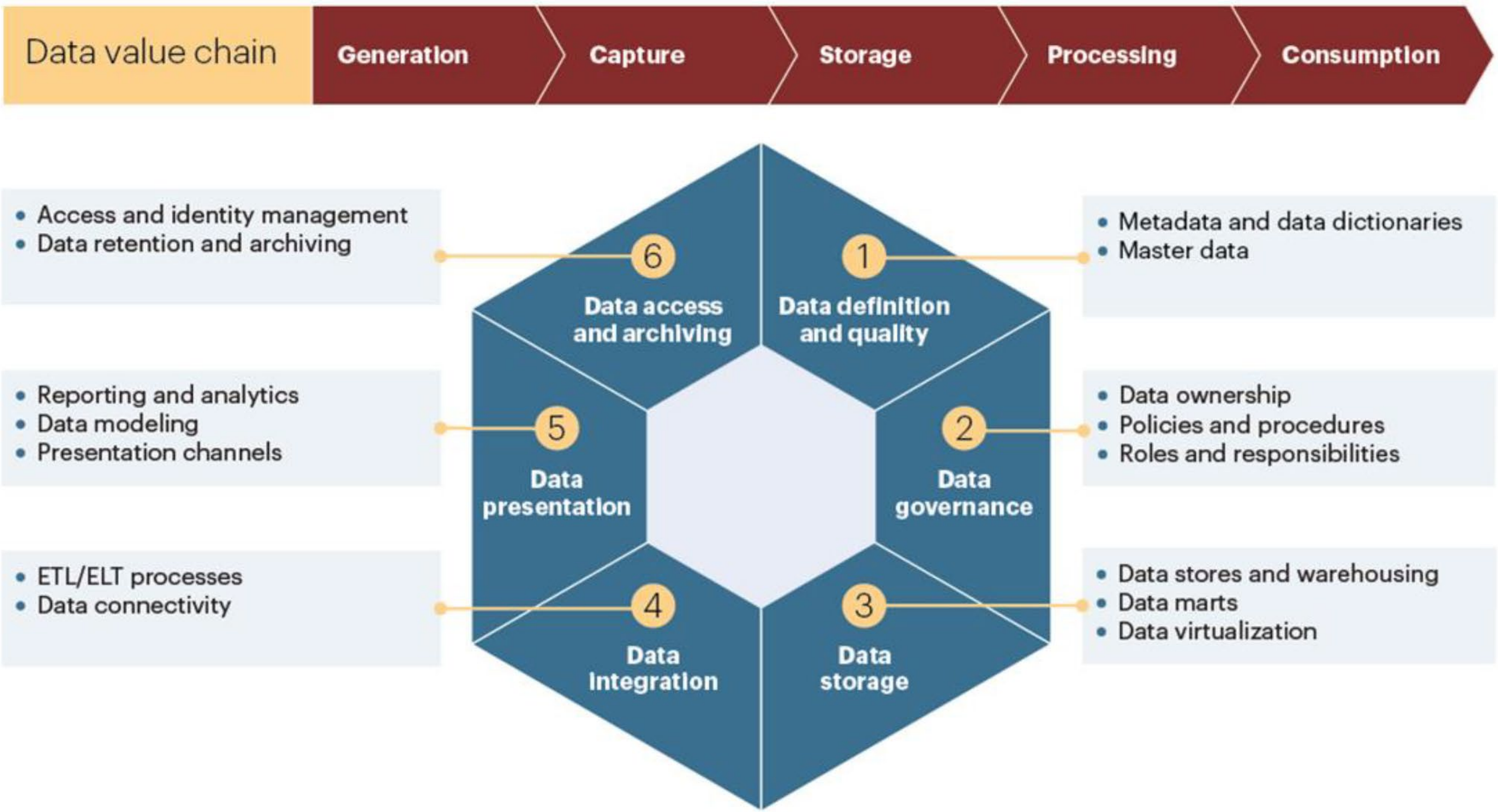


Architectural Patterns for Data Pipeline



The Dimensions of a Data Management Assessment

The dimensions of a data management assessment



Data value chain: Generation → Capture → Storage → Processing → Consumption

1. **Data definition and quality**
   - Metadata and data dictionaries
   - Master data
2. **Data governance**
   - Data ownership
   - Policies and procedures
   - Roles and responsibilities
3. **Data storage**
   - Data stores and warehousing
   - Data marts
   - Data virtualization
4. **Data integration**
   - ETL/ELT processes
   - Data connectivity
5. **Data presentation**
   - Reporting and analytics
   - Data modeling
   - Presentation channels
6. **Data access and archiving**
   - Access and identity management
   - Data retention and archiving

Source: A.T. Kearney analysis

Key Software Components of a DataOps Platform

> Data Pipeline Orchestration

> Testing and Production Quality

> Deployment Automation

> Data Science Model Deployment and Sandbox Management

> Data Virtualization, Versioning, and Test Data Management

>Data Integration and Unification

Who Are the Players

DATA KITCHEN
DATAOPS
inter|ana.
Nexla
Qubole
TRIFACTA
unravel

Composable Analytics
(https://composable.ai/)

Data Kitchen
(https://www.datakitchen.io/)

DataOps
(http://www.dataopsolutions.com/)

Interana
(https://www.interana.com/)

Nexla
(https://www.nexla.com/)

Qubole
(https://www.qubole.com/)

Trifacta
(https://www.trifacta.com/)

Unravel Data
(https://unraveldata.com/)

Comparison

Solution	Data Pipeline Orchestration	Testing and Production Quality	Automatic Deployment	Sandbox Management	Versioning	API	Command Line Interface	Cloud Support
Composable Analytics	✓	✓	✓		✓	✓		AWS, Azure
DataKitchen	✓	✓	✓		✓	✓	✓	AWS
DataOps	✓	✓						AWS
Interana		✓	✓	✓		✓	✓	AWS, Azure
Nexla	✓		✓		✓	✓		AWS, Google
Qubole		✓	✓	✓	✓	✓	✓	AWS, Azure, Google
Trifacta	✓	✓	✓	✓	✓	✓	✓	AWS, Azure, Google
Unravel Data	✓	✓	✓			✓	✓	AWS, Azure

References

> DataOps Explained: A Remedy For Ailing Data Pipelines:
https://www.eckerson.com/articles/dataops-explained-a-remedy-for-ailing-data-pipelines

> Gartner Hype Cycle for Data Management Positions Three Technologies in the Innovation Trigger Phase in 2018:
https://www.gartner.com/en/newsroom/press-releases/2018-09-11-gartner-hype-cycle-for-data-management-positions-three- technologies-in-the-innovation-trigger-phase-in-2018

> The Power of DataOps:
https://www.delphix.com/blog/power-dataops

> DataOps is NOT Just DevOps for Data: https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7

> The New Role of DataOps in Modern Organizations:
https://www.linkedin.com/pulse/new-role-dataops-modern-organizations-steven-wastie/

> The Power of DataOps:
https://www.cio.com/article/3236201/analytics/the-power-of-dataops.html

> Data Lifecycle Management:
https://www.oracle.com/a/ocom/docs/oracle-ds-data-ops-map-r.pdf

> High-Velocity Data Analytics with DataOps:
https://www.datakitchen.io/content/DataOpsWhitePaper.pdf

...it’s all about

DATA LEADERSHIP

DataOps Technical Components

> Test automation

> Code repositories

> Orchestration frameworks

> Collaboration and workflow management

> Metadata management

> Lineage and impact analysis

> Database management systems

> Data integration, preparation, and automation tools

> Analytics and visualization tools

> Monitoring and performance intelligence platforms

DataOps Manifest

- > Continually satisfy your customer
- > Value working analytics
- > Embrace change
- > It's a team sport
- > Daily interactions
- > Self-organize
- > Reduce heroism
- > Reflect
- > Analytics is code
- > Orchestrate
- > Make it reproducible
- > Disposable environments
- > Simplicity
- > Analytics is manufacturing
- > Quality is paramount
- > Monitor quality and performance
- > Reuse
- > Improve cycle times

Categories of DataOps Tools

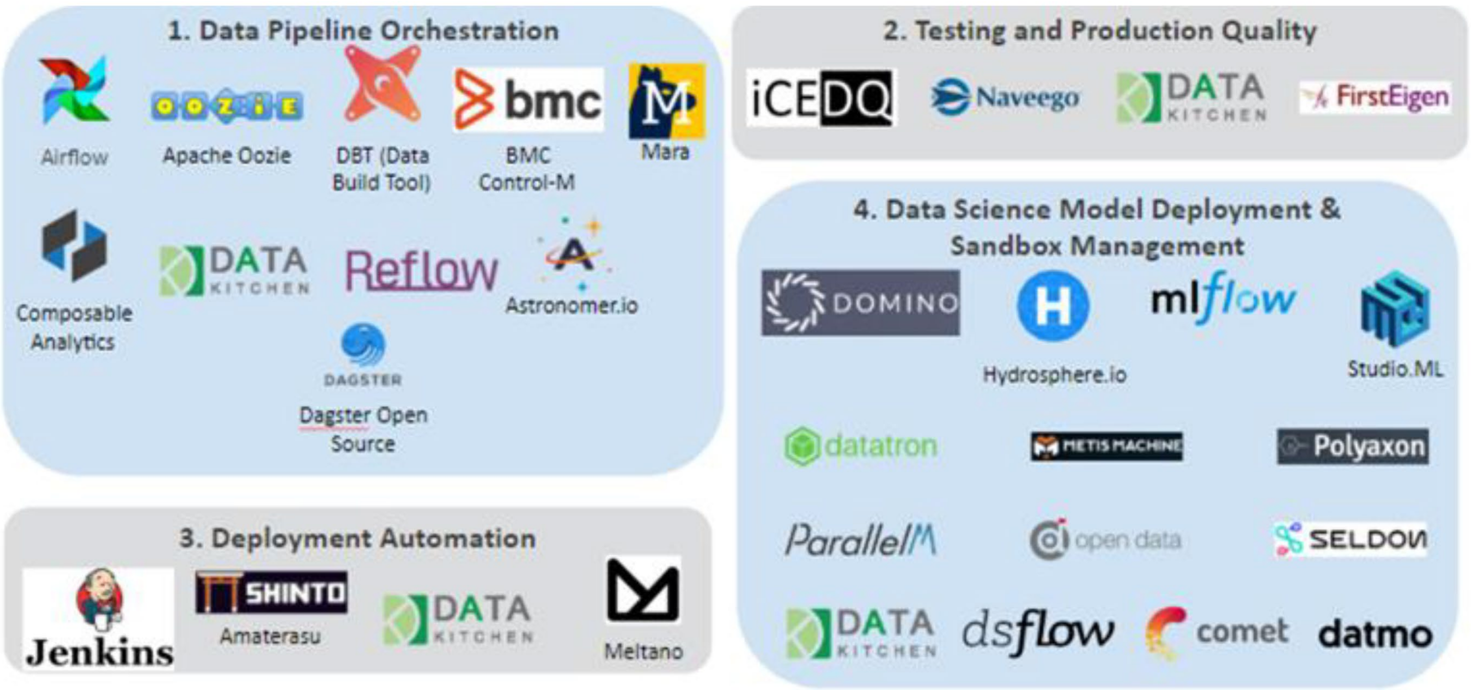
> **Orchestration and Operations Platforms.**
There are orchestration tools that automate the flow of of data and code multiple tools, across applications, and people. They act as a digital control room where all data sources and processes are managed and tuned. This reduces the complexity of managing complex data pipelines in a heterogeneous environment. Often these tools go beyond pre-production and provide monitoring capabilities to increase the visibility of performance and resource consumption for all stakeholders.

> **Data Warehouse Automation(DWA).**
These metadata-driven tool senable the automatic generation and deployment of data structures in a data warehouse, including staging areas, target databases, BI databases, and documentation. They are ideal for accelerating change management requests. Some DWA vendors are now extending their products to work with big data (Hadoop) and the cloud and handle more generic data-centric design, testing, and operations workflows.

> **Self-Service Data Preparation.**
These business-centric tools enable data analysts and other business users to build their own data and analytics pipelines, so they are no longer dependent upon the IT department. These tools facilitate a handoff from corporate IT, which uses data integration and extract, transform, and load (ETL) tools to ingest, clean, and lightly integrate data, while data analysts take the IT output and use data preparation tools manipulate the data to support local or individual use cases.

> **Data Science Platforms.**
Data science platforms are designed to accelerate, integrate, and automate the entire data science lifecycle, from data preparation and model creation to model deployment, monitoring, and management. Some platforms focus more on model development, others on model deployment, while some tackle the entire lifecycle.

DataOps Ecosystem



1. Data Pipeline Orchestration

- Airflow
- Apache Oozie
- DBT (Data Build Tool)
- BMC Control-M
- Mara
- Composable Analytics
- DATA KITCHEN
- Reflow
- Astronomer.io
- Dagster Open Source

2. Testing and Production Quality

- iCEDQ
- Naveego
- DATA KITCHEN
- FirstEigen

3. Deployment Automation

- Jenkins
- Amaterasu
- DATA KITCHEN
- Meltano

4. Data Science Model Deployment & Sandbox Management

- DOMINO
- Hydrosphere.io
- mlflow
- Studio.ML
- datatron
- METIS MACHINE
- Polyaxon
- ParallelM
- open data
- SELDON
- DATA KITCHEN
- dsflow
- comet
- datmo

aws | Partner Network | Morris & Opazo