THE DEVELOPER'S CONFERENCE

Trilha – BigData

DataOps: Estendendo as práticas de DevOps para BigData

Eduardo Hahn

DataLakers founder & DataOps Enthusiastic

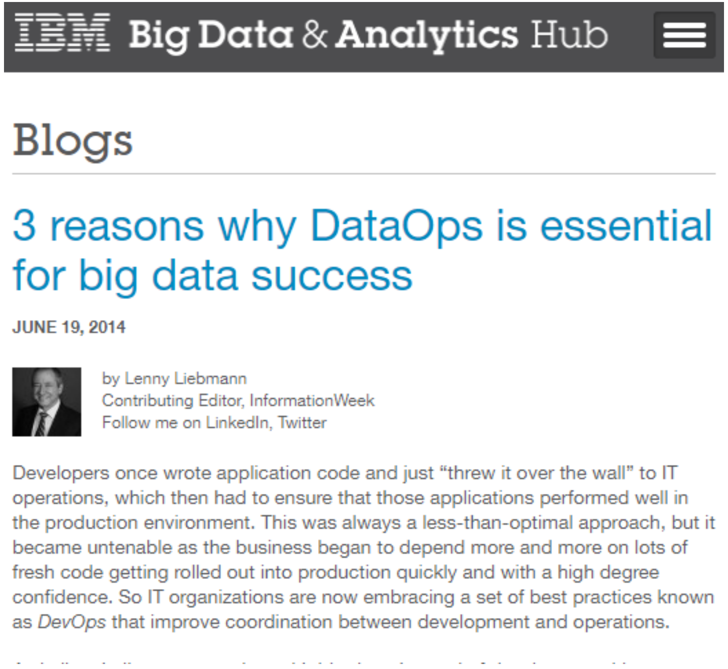
DataLakers
The Big Data Company

- + BuzzWord
- “**DataOps** is an automated, process-oriented methodology, used by analytic and data teams, to improve the quality and reduce the cycle time of data analytics.” Wikipedia
- “**DataOps** is about more than speed and quality. With a culture of continuous improvement, organizations can deliver data analytics solutions more efficiently, releasing valuable team members for more valuable activities, such as building innovative new products.” Eckerson Group

THE DEVELOPER'S CONFERENCE

Começo.....

- 2014....



IBM Big Data & Analytics Hub

Blogs

3 reasons why DataOps is essential for big data success

JUNE 19, 2014

by Lenny Liebmann
Contributing Editor, InformationWeek
Follow me on LinkedIn, Twitter

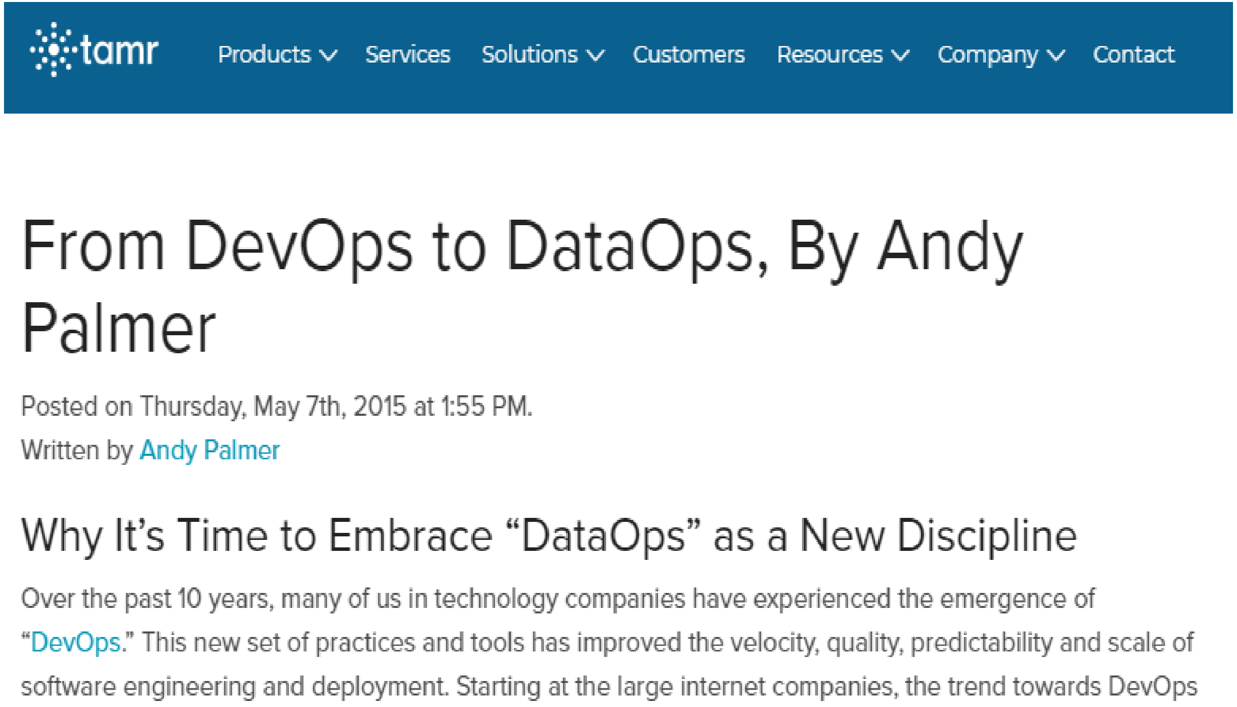
Developers once wrote application code and just "threw it over the wall" to IT operations, which then had to ensure that those applications performed well in the production environment. This was always a less-than-optimal approach, but it became untenable as the business began to depend more and more on lots of fresh code getting rolled out into production quickly and with a high degree confidence. So IT organizations are now embracing a set of best practices known as *DevOps* that improve coordination between development and operations.

"A ciência de dados é uma disciplina excepcionalmente importante hoje em dia. Mas essa ciência só é útil na medida em que pode ser executada de forma eficiente e confiável. E para que isso aconteça, você precisa de DataOps."

DataLakers
The Big Data Company

Começo.....

- 2015...



From DevOps to DataOps, By Andy Palmer

Posted on Thursday, May 7th, 2015 at 1:55 PM.

Written by Andy Palmer

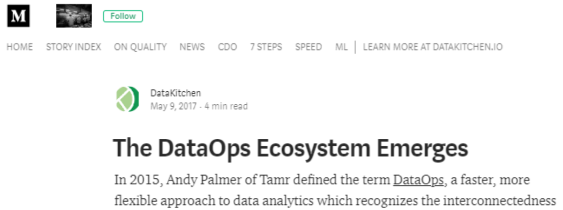
Why It's Time to Embrace "DataOps" as a New Discipline

Over the past 10 years, many of us in technology companies have experienced the emergence of "DevOps." This new set of practices and tools has improved the velocity, quality, predictability and scale of software engineering and deployment. Starting at the large internet companies, the trend towards DevOps

Acelerou.....

THE DEVELOPER'S CONFERENCE

- 2017....

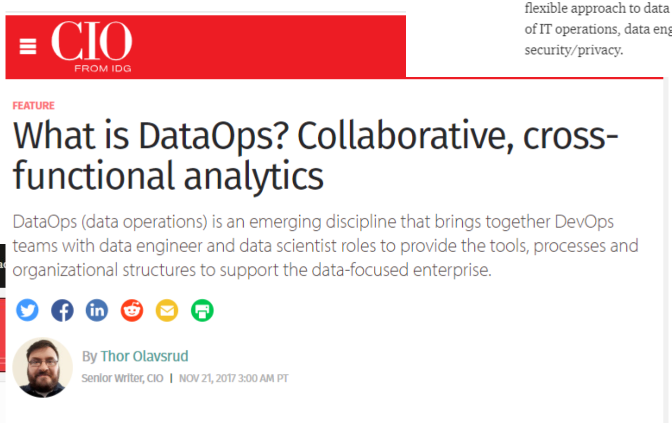

M Follow

HOME STORY INDEX ON QUALITY NEWS CDO 7 STEPS SPEED ML | LEARN MORE AT DATAKITCHEN.IO

DataKitchen
May 9, 2017 · 4 min read

The DataOps Ecosystem Emerges

In 2015, Andy Palmer of Tamr defined the term DataOps, a faster, more flexible approach to data analytics which recognizes the interconnectedness of IT operations, data engineering, data integration, data quality and data security/privacy.


CIO FROM IDG

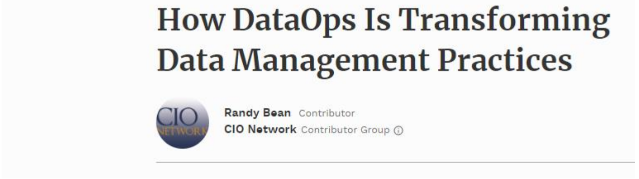
FEATURE

What is DataOps? Collaborative, cross-functional analytics

DataOps (data operations) is an emerging discipline that brings together DevOps teams with data engineer and data scientist roles to provide the tools, processes and organizational structures to support the data-focused enterprise.

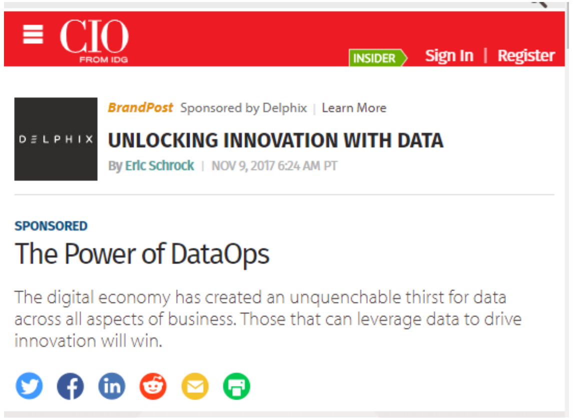
By Thor Olavsrud
Senior Writer, CIO | NOV 21, 2017 3:00 AM PT

Forbes Billionaires Innovation Lea


3,992 views | Apr 11, 2018, 01:38pm

How DataOps Is Transforming Data Management Practices

Randy Bean Contributor
CIO Network Contributor Group


CIO FROM IDG INSIDER Sign In | Register

BrandPost Sponsored by Delphix | Learn More
DELPHIX
UNLOCKING INNOVATION WITH DATA
By Eric Schrock | NOV 9, 2017 6:24 AM PT

SPONSORED

The Power of DataOps

The digital economy has created an unquenchable thirst for data across all aspects of business. Those that can leverage data to drive innovation will win.

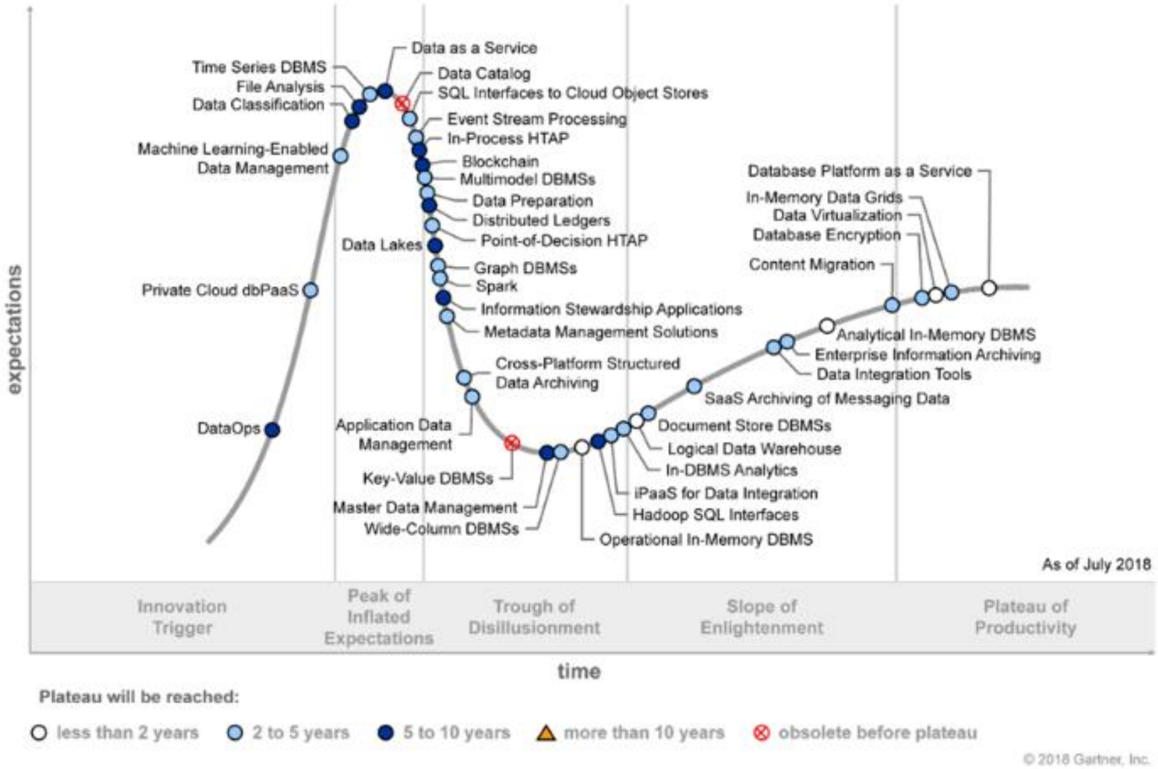
DataLakers
The Big Data Company

THE DEVELOPER'S CONFERENCE

Agora sim.....

- 2018...

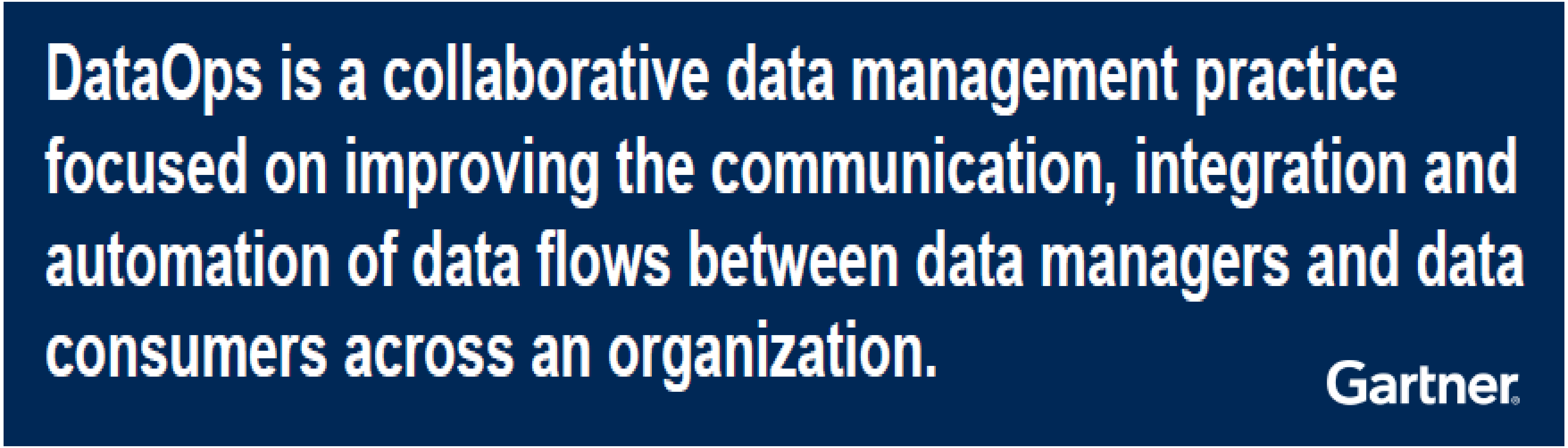
Figure 1. Hype Cycle for Data Management, 2018



© 2018 Gartner, Inc.

DataLakers
The Big Data Company

Agora sim.....

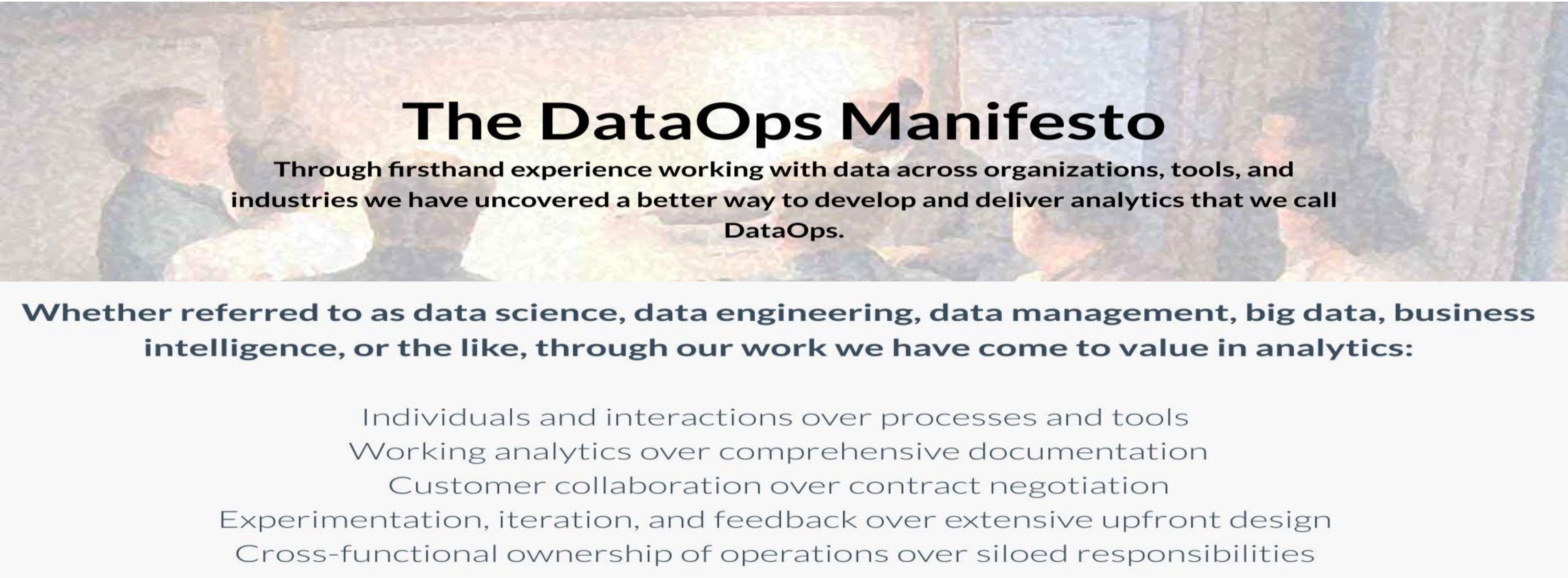


DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization.

Gartner.

DataOps Manifesto

THE
DEVELOPER'S
CONFERENCE



The DataOps Manifesto

Through firsthand experience working with data across organizations, tools, and industries we have uncovered a better way to develop and deliver analytics that we call DataOps.

Whether referred to as data science, data engineering, data management, big data, business intelligence, or the like, through our work we have come to value in analytics:

Individuals and interactions over processes and tools
Working analytics over comprehensive documentation
Customer collaboration over contract negotiation
Experimentation, iteration, and feedback over extensive upfront design
Cross-functional ownership of operations over siloed responsibilities

dataopsmanifesto.org/

DataLakers
The Big Data Company

DataOps Manifesto

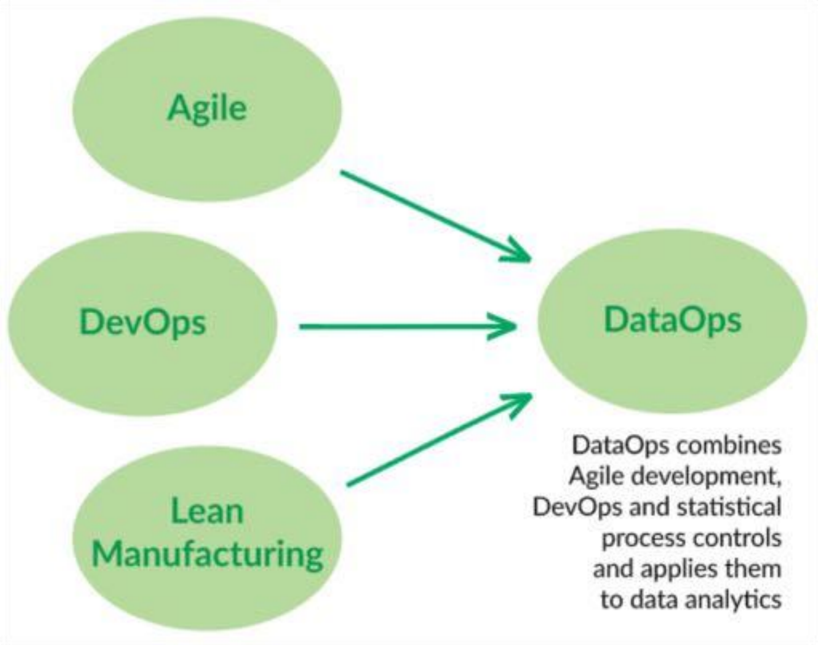
- Princípios do DataOps
  - 1. Satisfaça continuamente o seu cliente
  - 2. Valor do trabalho analítico
  - 3. Abrace a mudança
  - 4. É um esporte em equipe
  - 5. Interações diárias
  - 6. Auto-organização
  - 7. Reduza o heroísmo
  - 8. Reflita
  - 9. Os códigos

- 10. Orquestração
- 11. Faça tudo ser reproduzível
- 12. Ambientes descartáveis
- 13. Simplicidade
- 14. Análise de dados é manufatura
- 15. A qualidade é primordial
- 16. Monitorar a qualidade e o desempenho
- 17. Reutilizar
- 18. Melhorar os tempos dos ciclos

Genesis of DataOps

People,
Process,
Organization

Technical
Environment

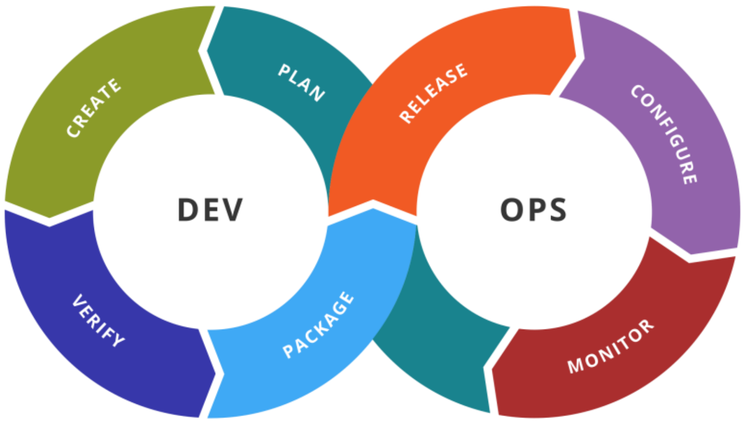


DataOps combines Agile development, DevOps and statistical process controls and applies them to data analytics

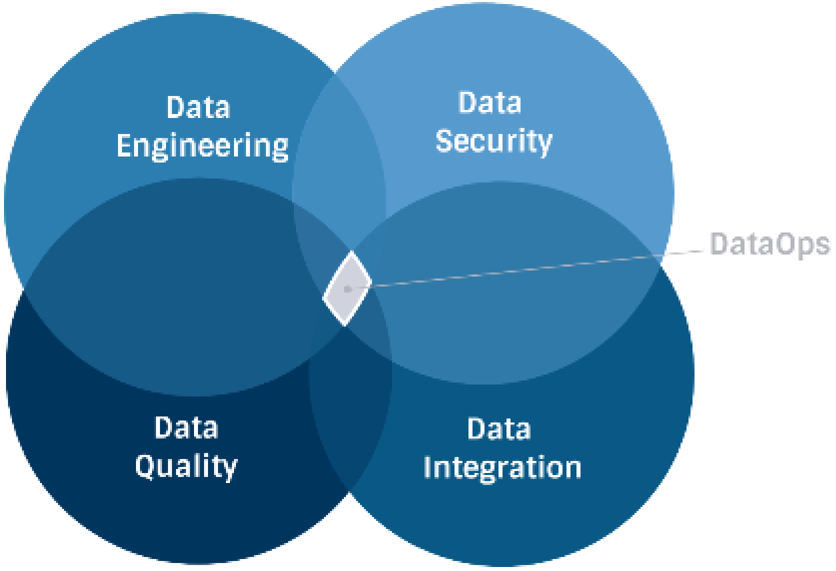
4 “As” de DataOps

- Automatize e monitorar pipelines
- Automatizar implantações
- Automatizar e monitorar a qualidade
- Automatizar sandbox

DataOps is NOT Just DevOps for Data

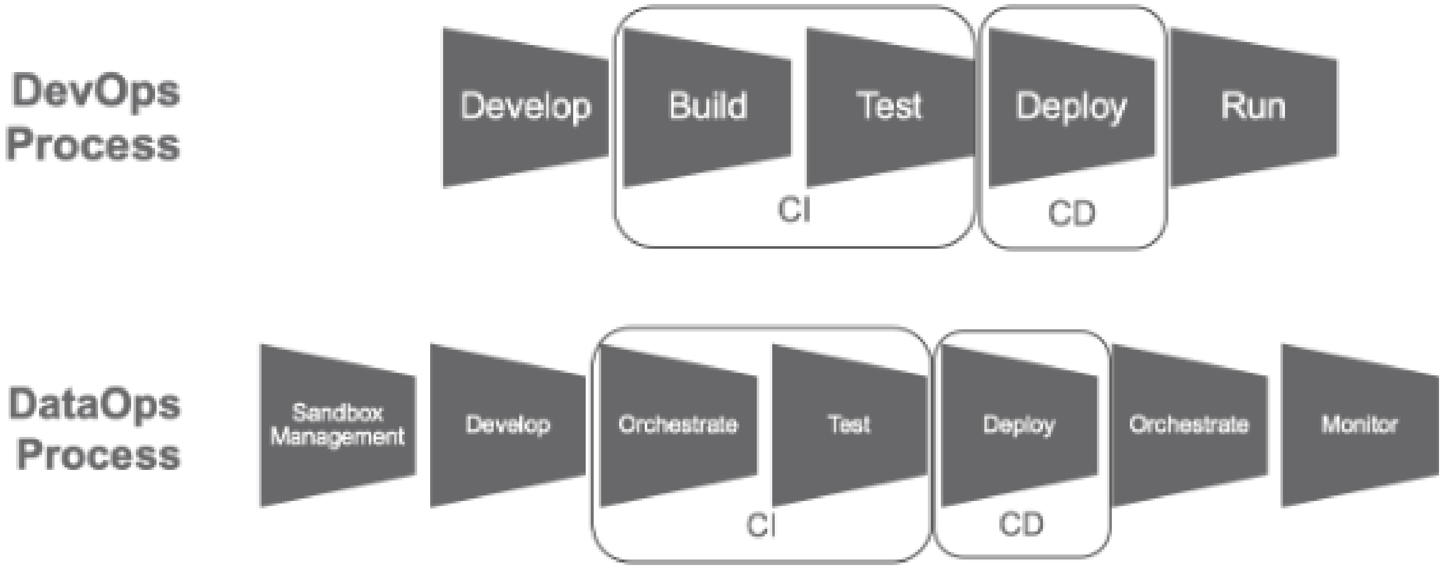


DEVOPS



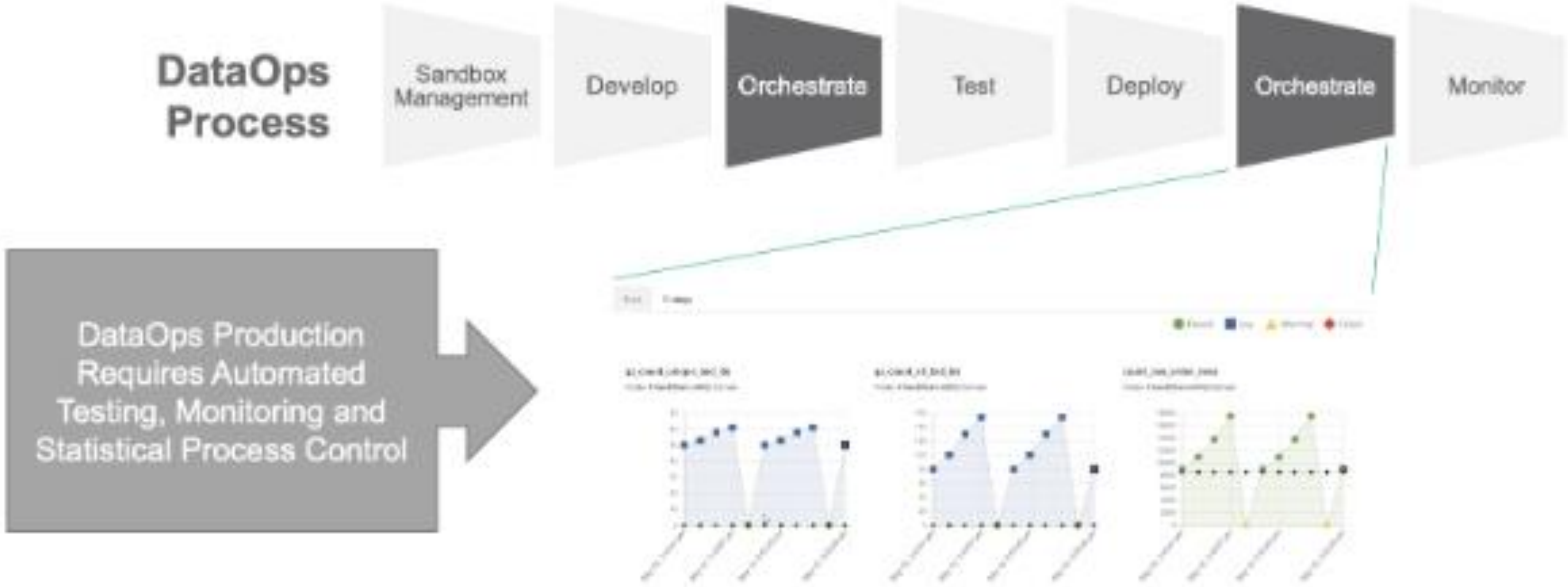
DATAOPS

DataOps is NOT Just DevOps for Data



DataOps is NOT Just DevOps for Data

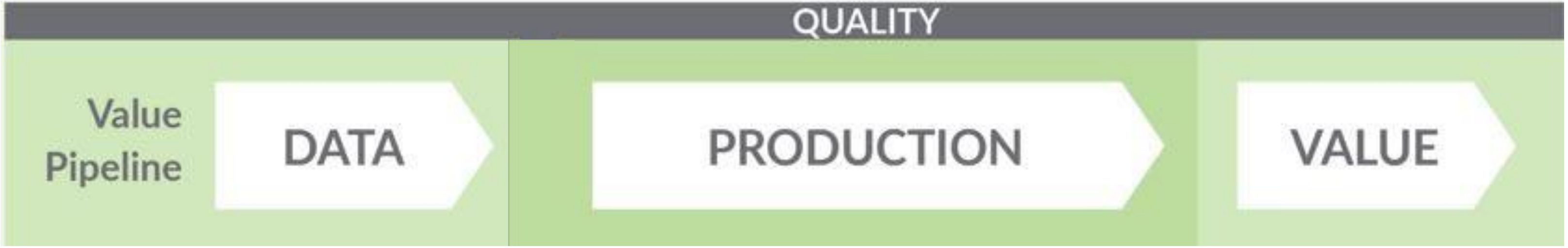
THE DEVELOPER'S CONFERENCE


DataOps Process
Sandbox Management
Develop
Orchestrate
Test
Deploy
Orchestrate
Monitor
DataOps Production Requires Automated Testing, Monitoring and Statistical Process Control

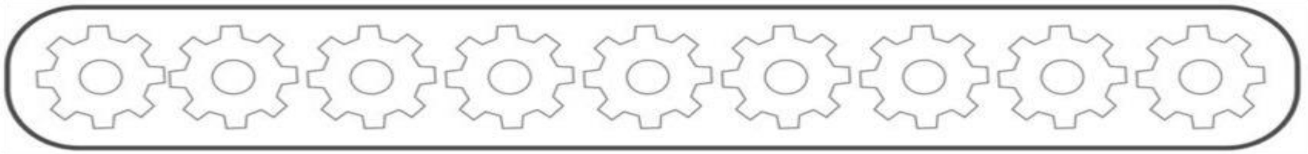
DataLakers
The Big Data Company

DataOps is NOT Just DevOps for Data

Automatize e monitorar pipelines

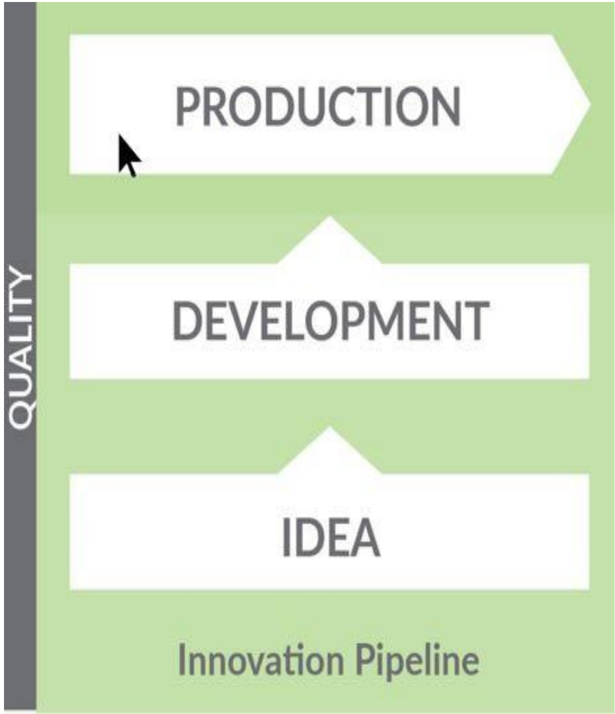


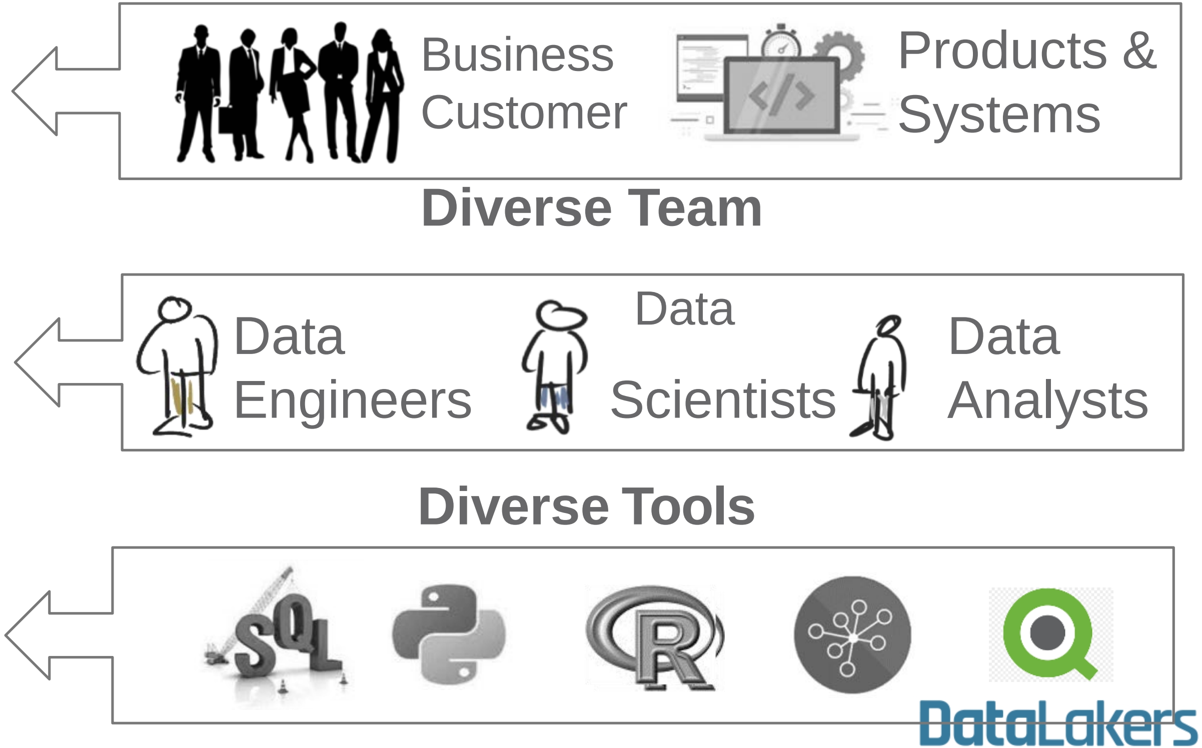
Access: Python Code → **Transform**: SQL Code, ETL → **Model**: R Code → **Visualize**: DataViz → **Report**:



DataOps is NOT Just DevOps for Data

Automatizar implantações





Business Customer

Products & Systems

Diverse Team

Data Engineers

Data Scientists

Data Analysts

Diverse Tools

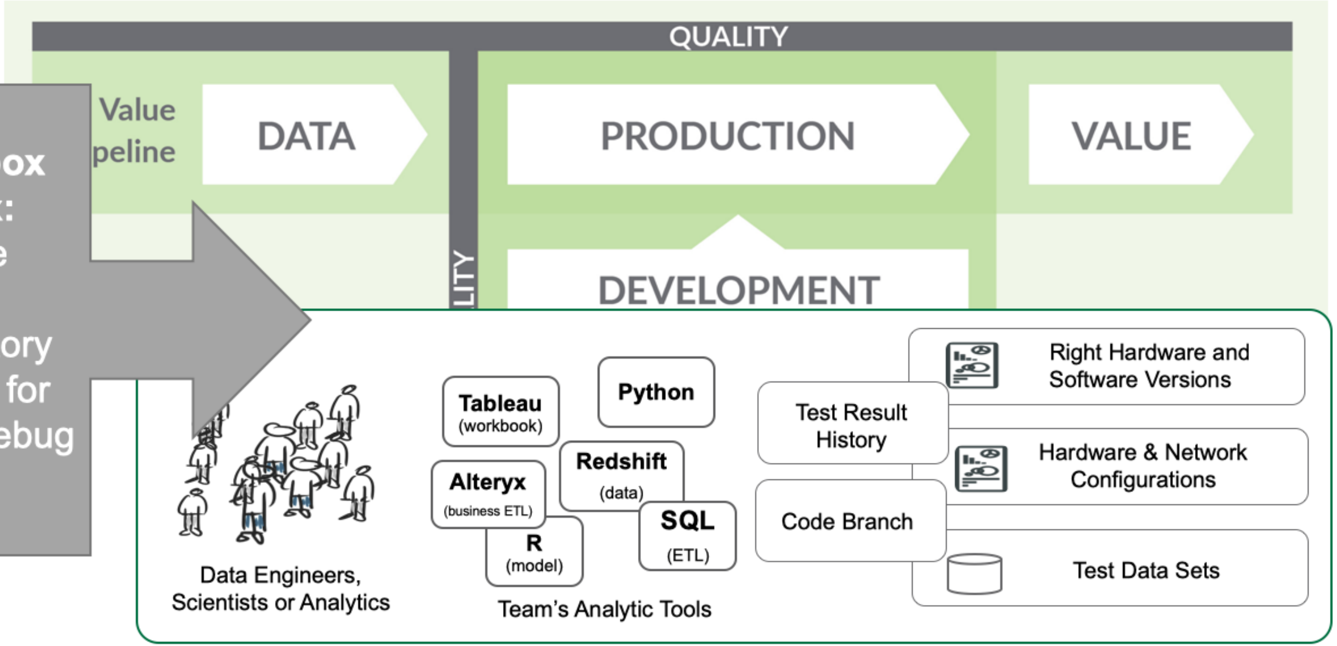
DataOps is NOT Just DevOps for Data

THE DEVELOPER'S CONFERENCE

Automatizar sandbox

Development Sandbox Creation is Complex:

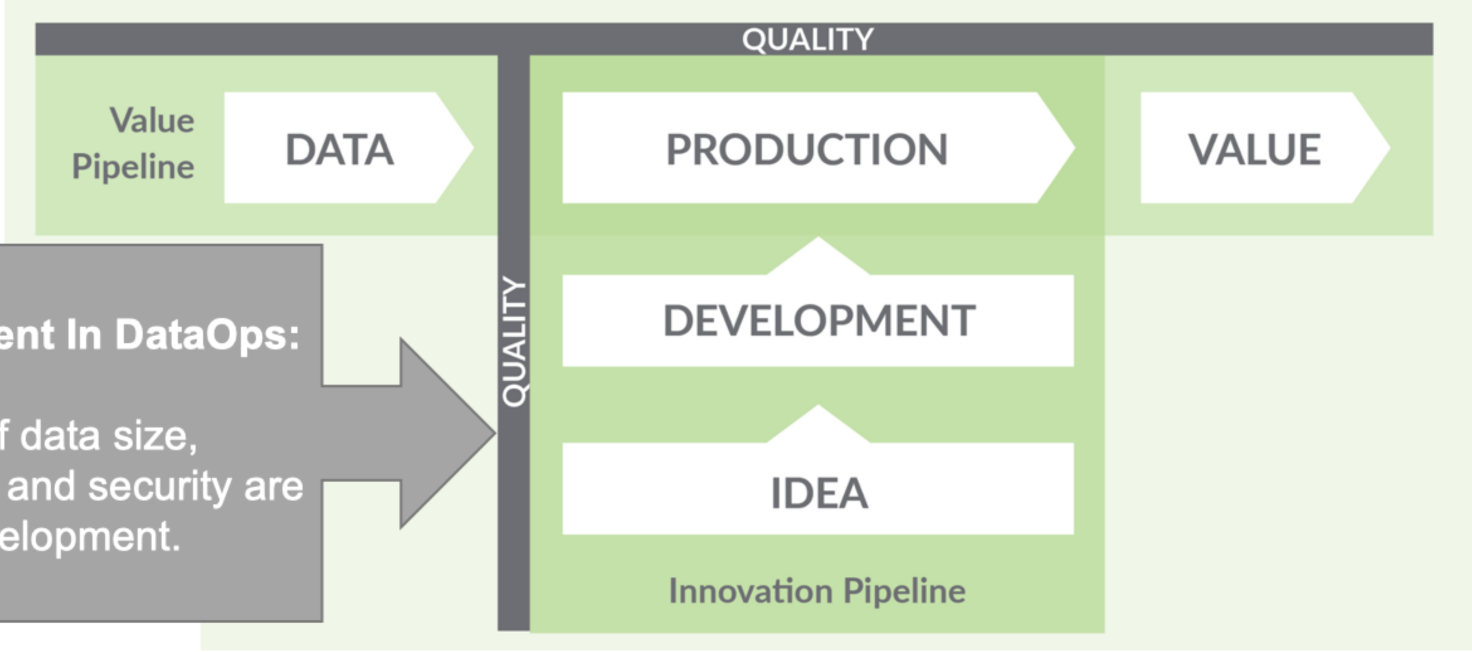
Hard to create the right set of data, tools, people, history and configuration for a fast build test debug cycle



DataLakers
The Big Data Company

DataOps is NOT Just DevOps for Data

Automatizar e monitorar a qualidade



Test Data Management In DataOps:

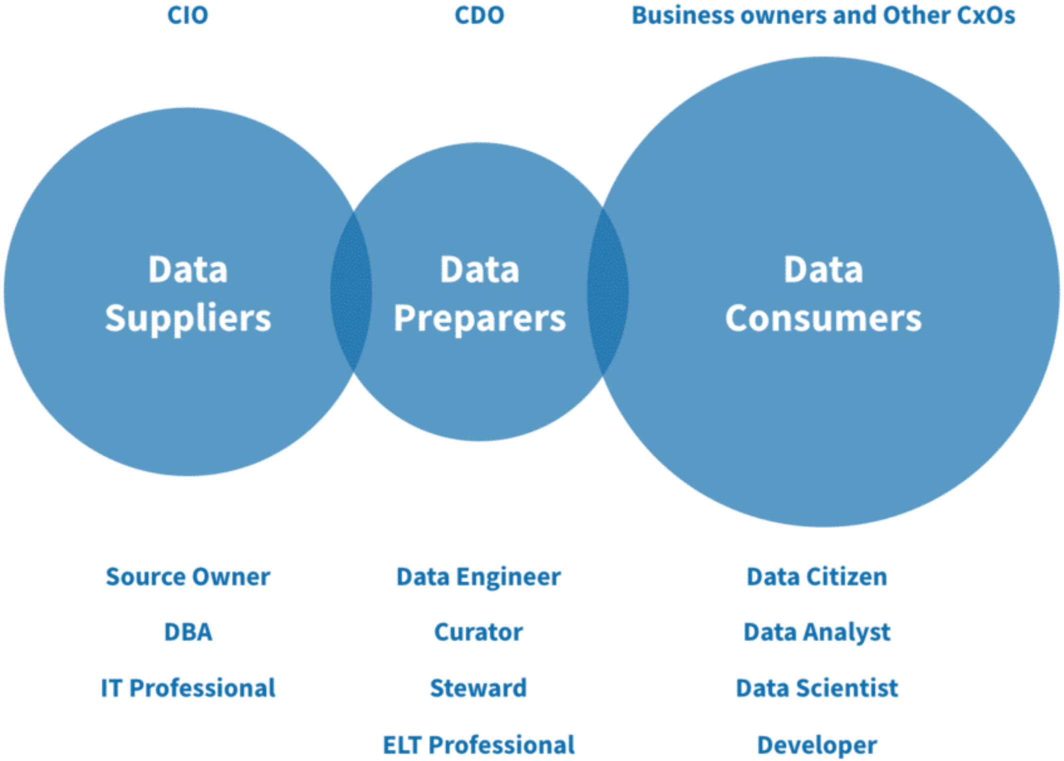
Consideration of data size, complexity, sampling and security are required in development.

THE DEVELOPER'S CONFERENCE

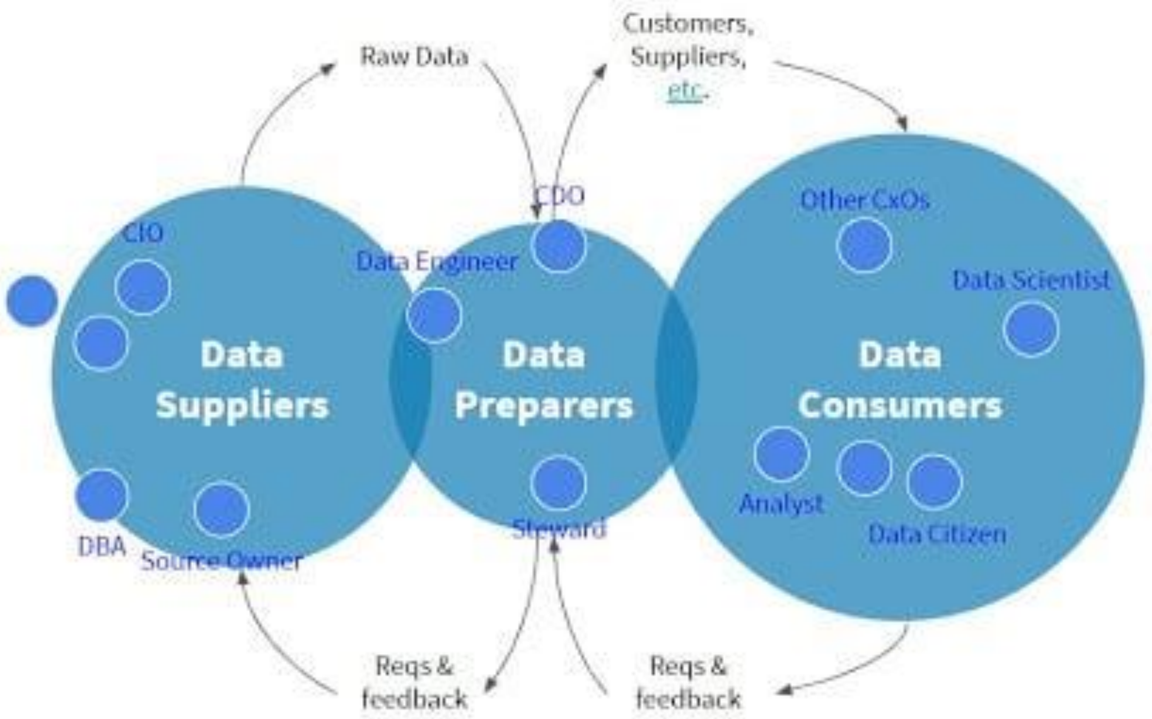
The People of DataOps

DataLakers
The Big Data Company

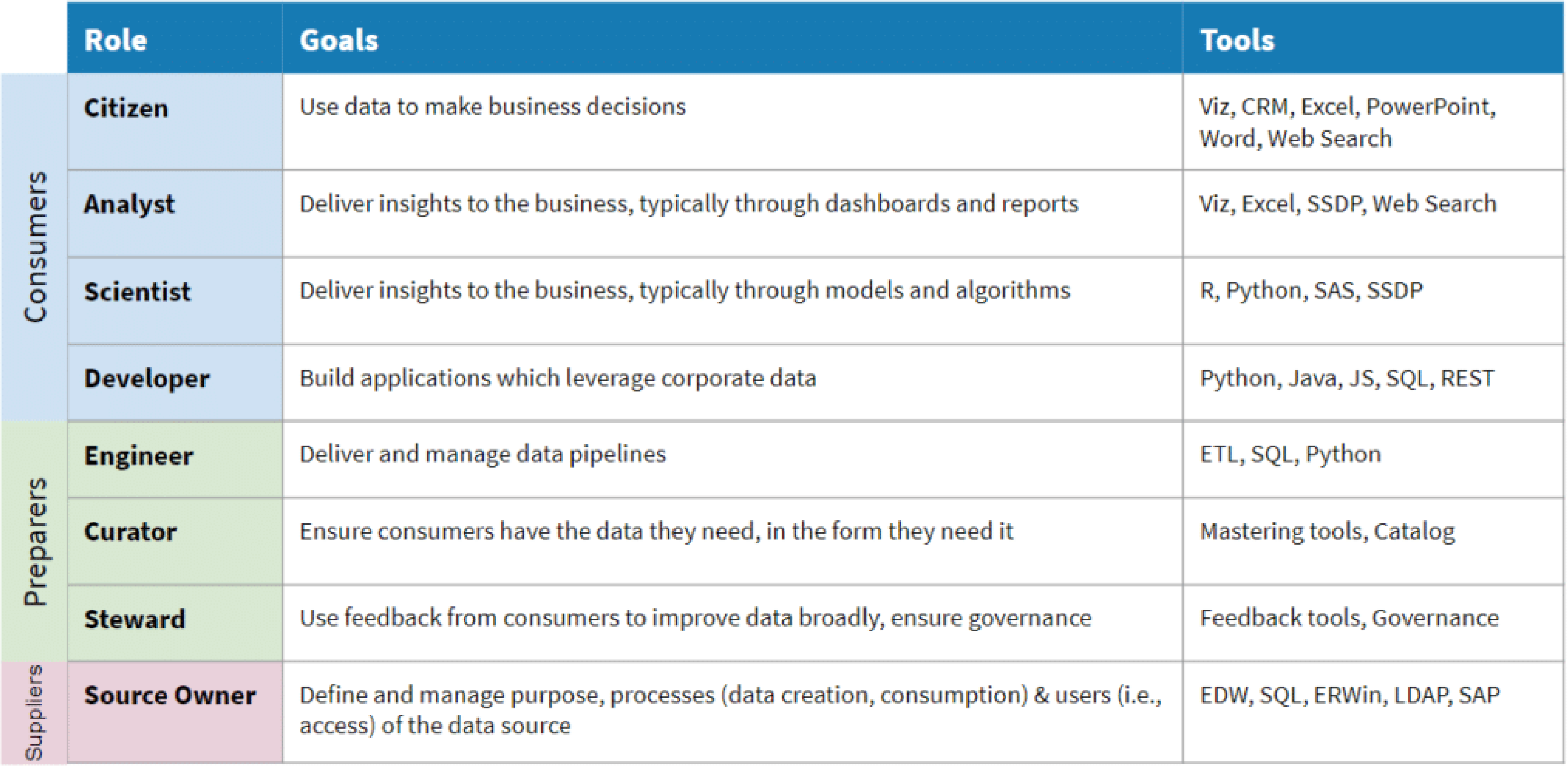
The People of DataOps



The People of DataOps



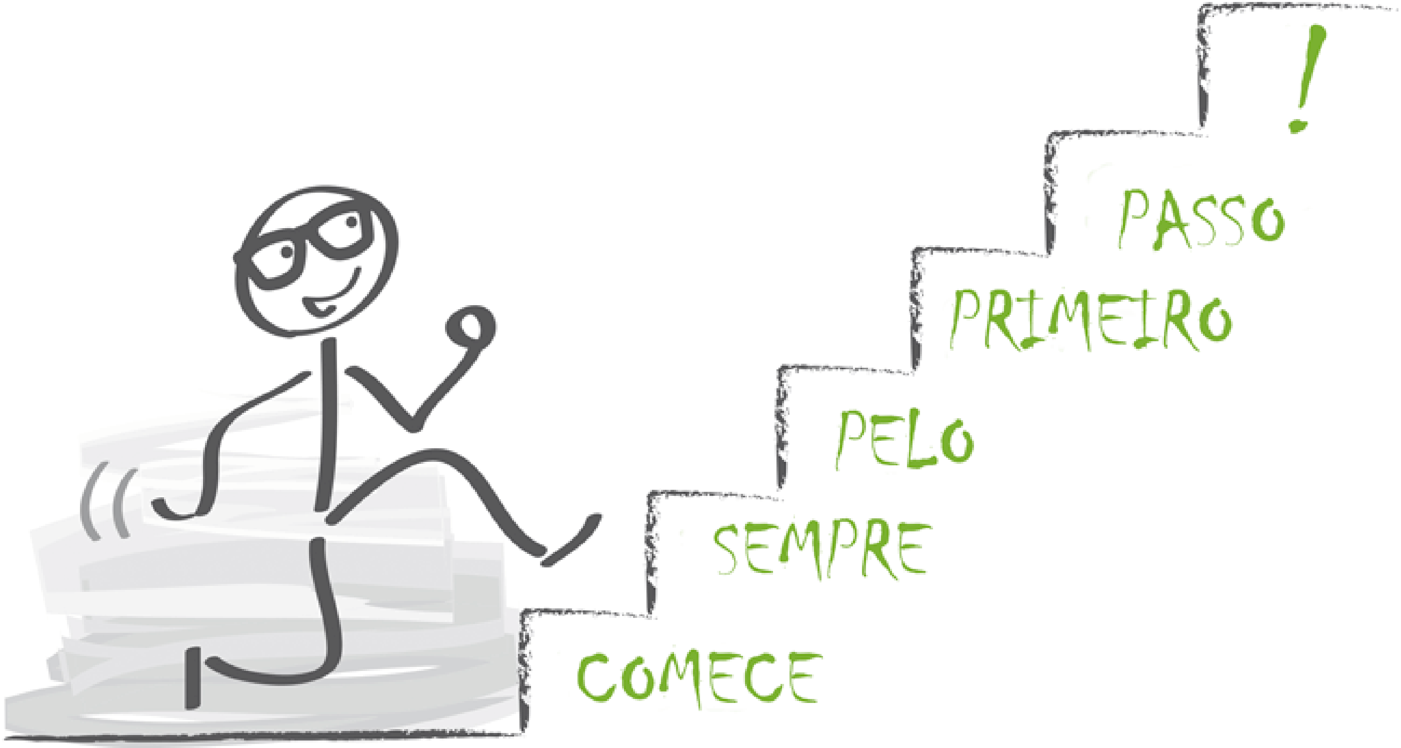
The People of DataOps



	Role	Goals	Tools
Consumers	Citizen	Use data to make business decisions	Viz, CRM, Excel, PowerPoint, Word, Web Search
	Analyst	Deliver insights to the business, typically through dashboards and reports	Viz, Excel, SSDP, Web Search
	Scientist	Deliver insights to the business, typically through models and algorithms	R, Python, SAS, SSDP
	Developer	Build applications which leverage corporate data	Python, Java, JS, SQL, REST
Preparers	Engineer	Deliver and manage data pipelines	ETL, SQL, Python
	Curator	Ensure consumers have the data they need, in the form they need it	Mastering tools, Catalog
	Steward	Use feedback from consumers to improve data broadly, ensure governance	Feedback tools, Governance
Suppliers	Source Owner	Define and manage purpose, processes (data creation, consumption) & users (i.e., access) of the data source	EDW, SQL, ERWin, LDAP, SAP

Passos para implementar DataOps

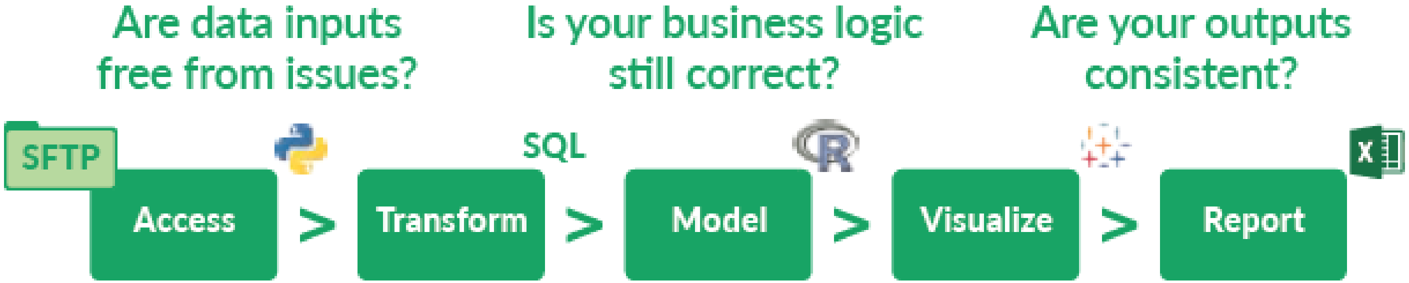
THE DEVELOPER'S CONFERENCE



DataLakers
The Big Data Company

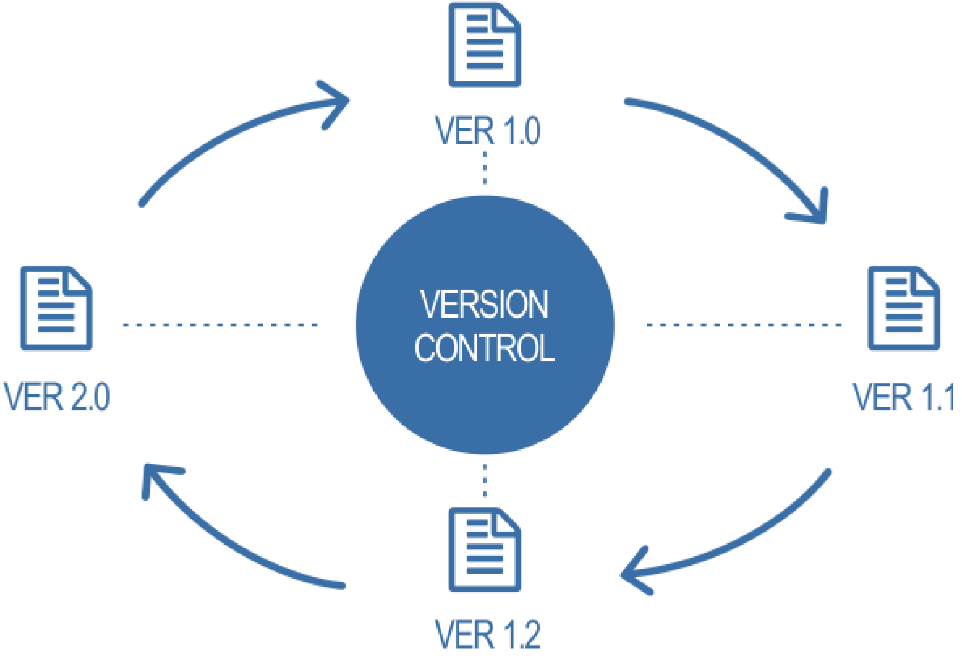
Passos para implementar DataOps

- Passo 1: implemente Data Test e Logic Test


Are data inputs free from issues?
Is your business logic still correct?
Are your outputs consistent?
SFTP
SQL
Access
>
Transform
>
Model
>
Visualize
>
Report

Passos para implementar DataOps

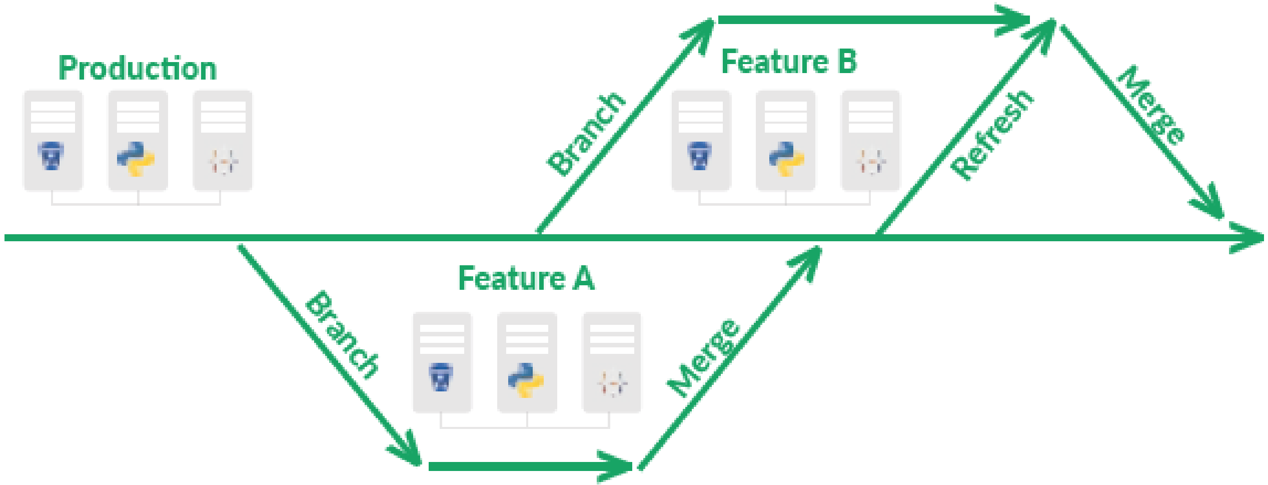
- Passo 2: use controle de versão (Branch&Merge).



The Big Data Company

Passos para implementar DataOps

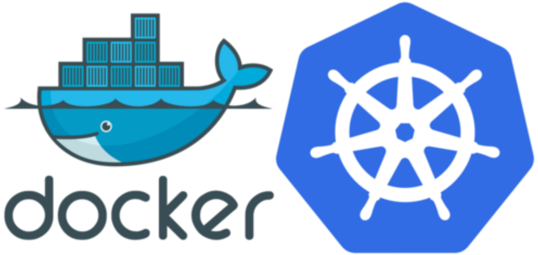
- Passo 3: Use múltiplos ambientes



Passos para implementar DataOps

Passo 4: Reuso e containers

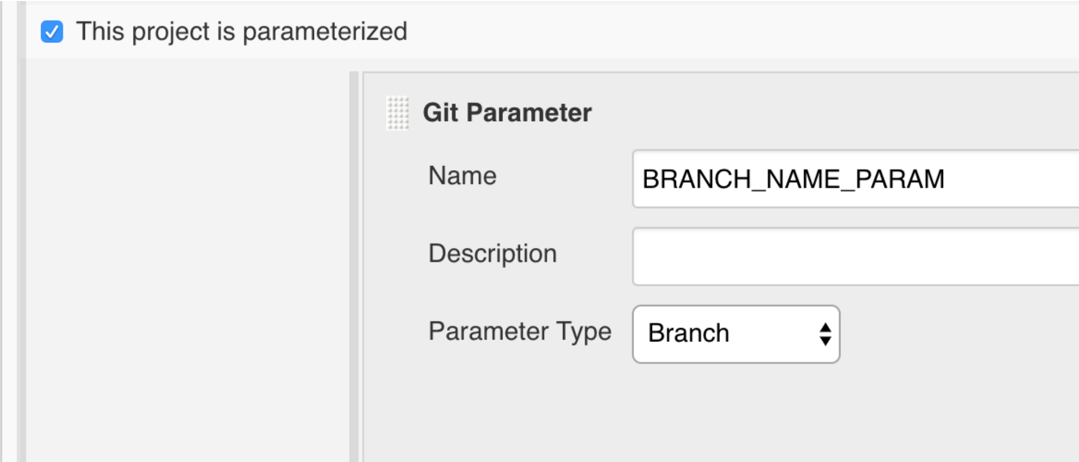
Outro método de aumento de produtividade para equipes é a capacidade de reutilizar e conter o código. Cada etapa intermediária no pipeline de análise de dados recebe a saída de um estágio anterior e fornece entrada para o próximo estágio. É mais fácil para os outros membros da equipe reutilizar componentes menores, se eles puderem ser segmentados ou conteinerização. Melhor caminho é usar Docker



Passos para implementar DataOps

THE DEVELOPER'S CONFERENCE

- Passo 5: Parametrize seus processos



☑ This project is parameterized

Git Parameter

Name	BRANCH_NAME_PARAM
Description	
Parameter Type	Branch

DataLakers
The Big Data Company

Passos para implementar DataOps

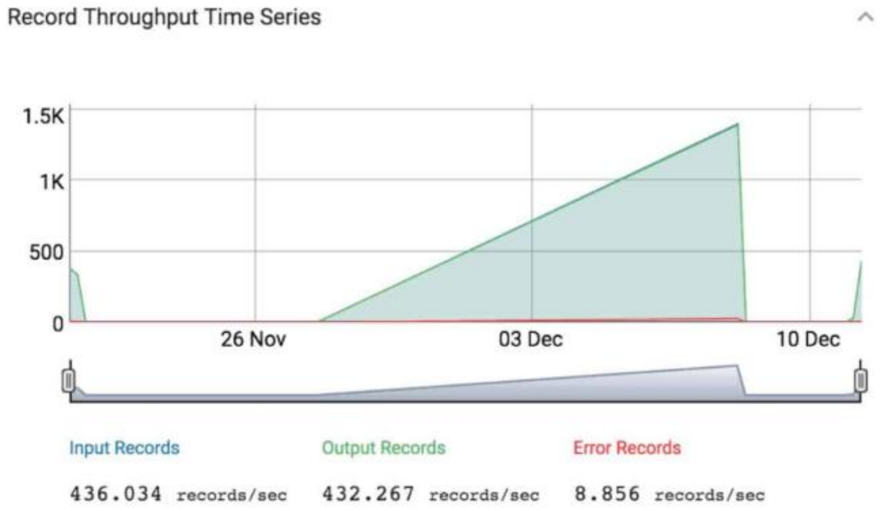
Passo 6: Use Simple Unique Storage

- Data Lake: os dados são movidos de diferentes silos de dados para um repositório comum, é muito mais fácil para uma equipe de análise de dados trabalhar com ele.

Passos para implementar DataOps

› Passo 7: Medir todas as etapas do pipeline

Defina métricas de ponta a ponta para sua arquitetura. Identifique pontos de melhorias e problemas de desempenho. Visualize uma arquitetura de dados online, visualizando como os sistemas evoluem.



Input Records 436.034 records/sec

Output Records 432.267 records/sec

Error Records 8.856 records/sec

DataOps Ecosystem

Platform Solutions

composable

DATA KITCHEN

tamr

StreamSets

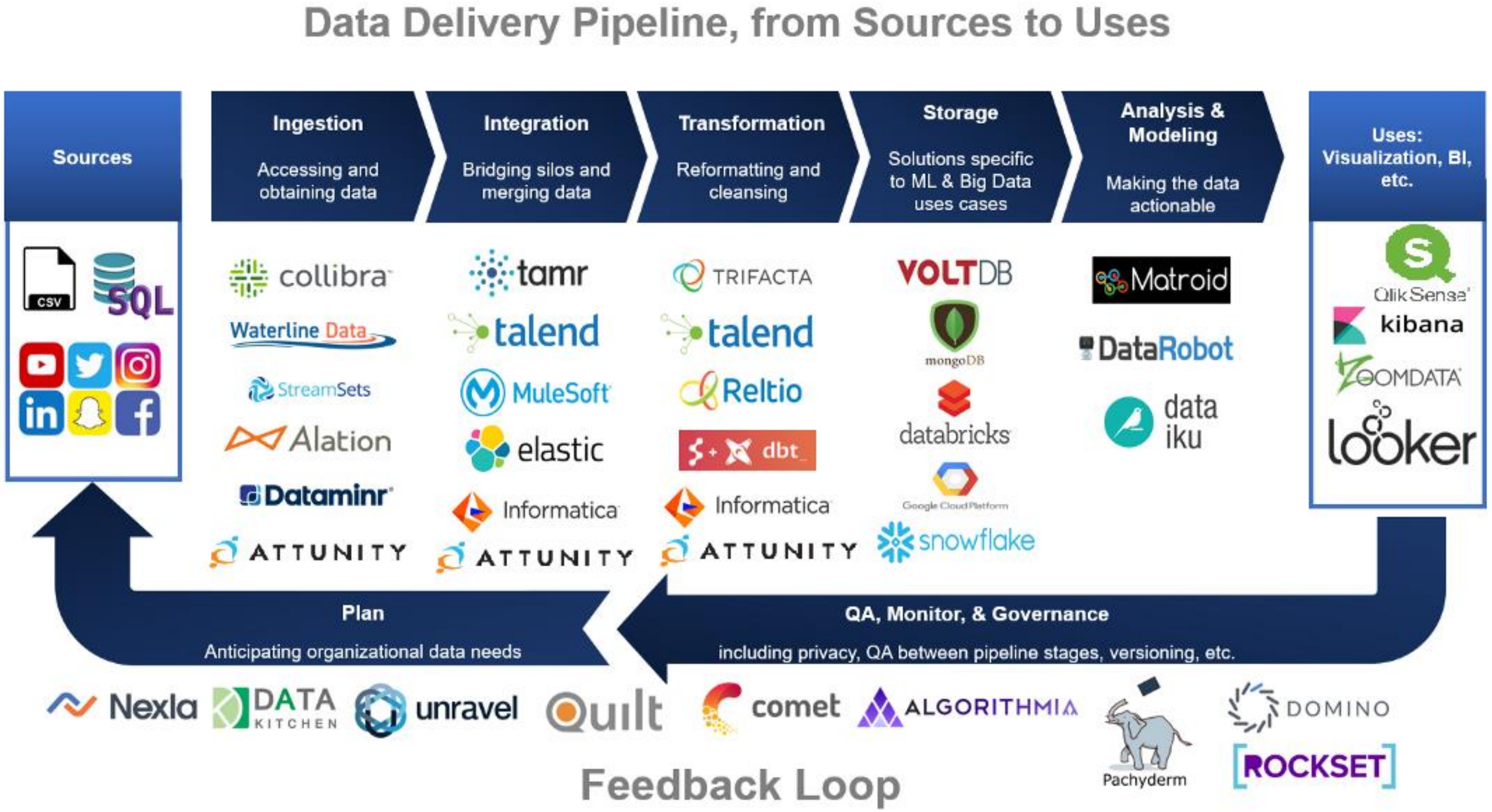
kinæsis
data / intelligence

talend

ATTUNITY

DataOps Ecosystem

THE DEVELOPER'S CONFERENCE



Data Delivery Pipeline, from Sources to Uses

Sources

Ingestion
Accessing and obtaining data

Integration
Bridging silos and merging data

Transformation
Reformatting and cleansing

Storage
Solutions specific to ML & Big Data uses cases

Analysis & Modeling
Making the data actionable

Uses: Visualization, BI, etc.

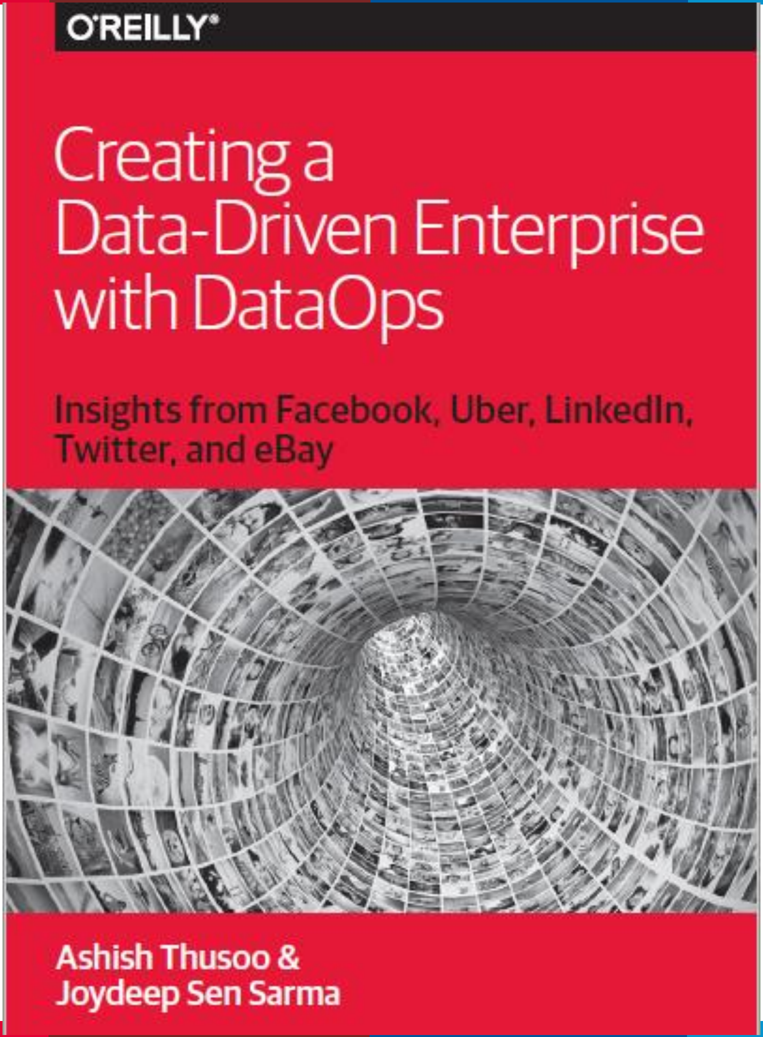
Plan
Anticipating organizational data needs

QA, Monitor, & Governance
including privacy, QA between pipeline stages, versioning, etc.

Feedback Loop

DataLakers
The Big Data Company

References



O'REILLY®

Creating a Data-Driven Enterprise with DataOps

Insights from Facebook, Uber, LinkedIn, Twitter, and eBay

Ashish Thusoo & Joydeep Sen Sarma

THE DEVELOPER'S CONFERENCE

DataLakers
The Big Data Company

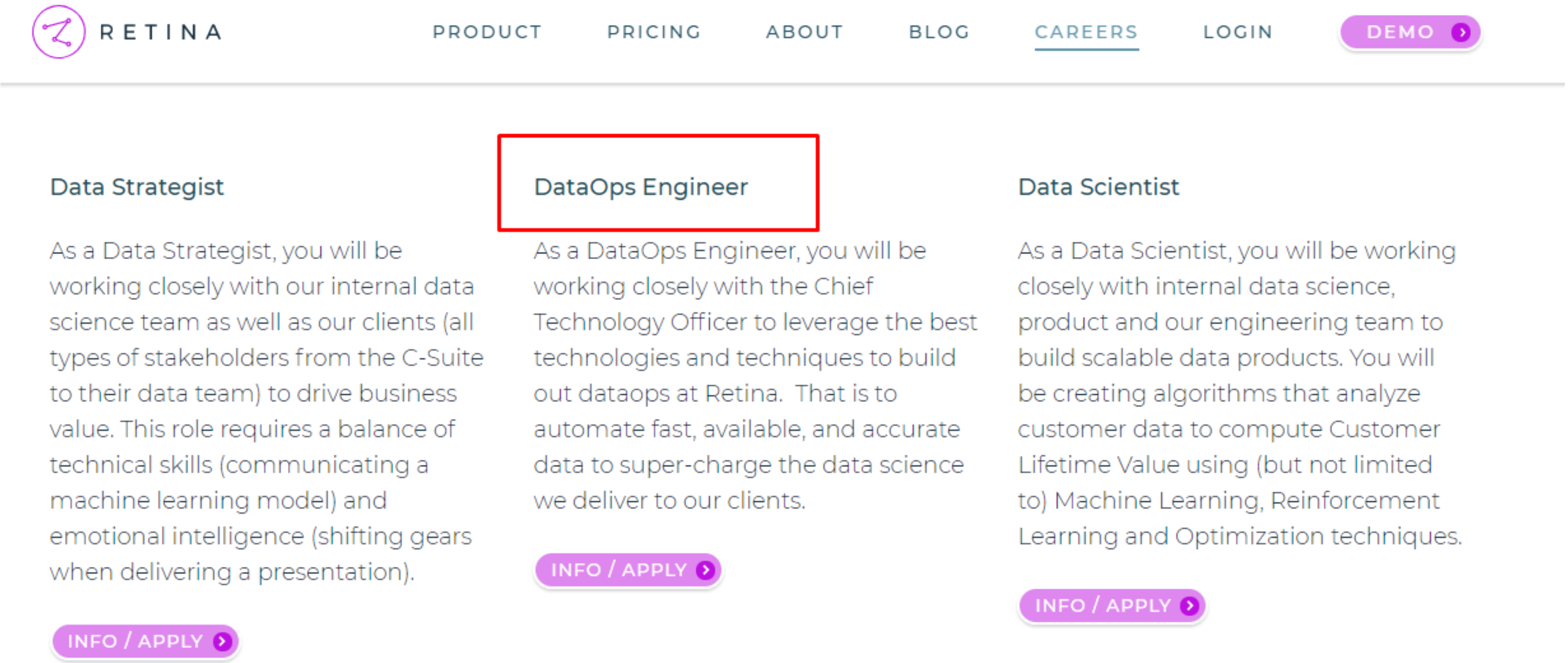
References

- **DataOps Ecosystem - //**medium.com/data-ops/2017-the-year-of-dataops-b2023c17d2af
- **DataOps for Government (State of Connecticut) -** //github.com/OpenDataCT/DataOps
- **Creating a Data-Driven Enterprise with DataOps - //**www.oreilly.com/data/free/creating-a-data-driven-enterprise-with-dataops.csp
- http://dataopsmanifesto.org/
- **DataOps—It's a Secret** - //www.datasciencecentral.com/profiles/blogs/dataops-it-s-a-secret
- **The Power of DataOps** - //www.delphix.com/blog/power-dataops
- **Building a DataOps Team** - //medium.com/data-ops/building-a-dataops-team-abc375e0a6bc
- **DataOps: Industrializing Data and Analytics -** //www.eckerson.com/articles/dataops-industrializing-data-and-analytics?content=dataops-industrializing-data-and-analytics

Concluindo....

- As empresas que desejam implementar DataOps devem concentrar seus esforços em três áreas:
  - Cultura
  - Processos
  - Tecnologia

Oportunidades estão chegando....



Data Strategist

As a Data Strategist, you will be working closely with our internal data science team as well as our clients (all types of stakeholders from the C-Suite to their data team) to drive business value. This role requires a balance of technical skills (communicating a machine learning model) and emotional intelligence (shifting gears when delivering a presentation).

DataOps Engineer

As a DataOps Engineer, you will be working closely with the Chief Technology Officer to leverage the best technologies and techniques to build out dataops at Retina. That is to automate fast, available, and accurate data to super-charge the data science we deliver to our clients.

Data Scientist

As a Data Scientist, you will be working closely with internal data science, product and our engineering team to build scalable data products. You will be creating algorithms that analyze customer data to compute Customer Lifetime Value using (but not limited to) Machine Learning, Reinforcement Learning and Optimization techniques.

About me

THE DEVELOPER'S CONFERENCE

Big Family

- Eduardo Hahn
  - Founder DataLakers Tecnologia
  - Data Lover & DataOps Enthusiastic
  - eduardo.hahn@datalakers.com.br
  - @eduardohahn
  - /in/eduardohahn3

Partners

cloudera talend ATTUNITY

DataLakers
The Big Data Company

THE DEVELOPER'S CONFERENCE

Trilha – BigData

DataOps: Estendendo as práticas de DevOps para BigData

Eduardo Hahn

DataLakers founder & DataOps Enthusiastic

DataLakers
The Big Data Company