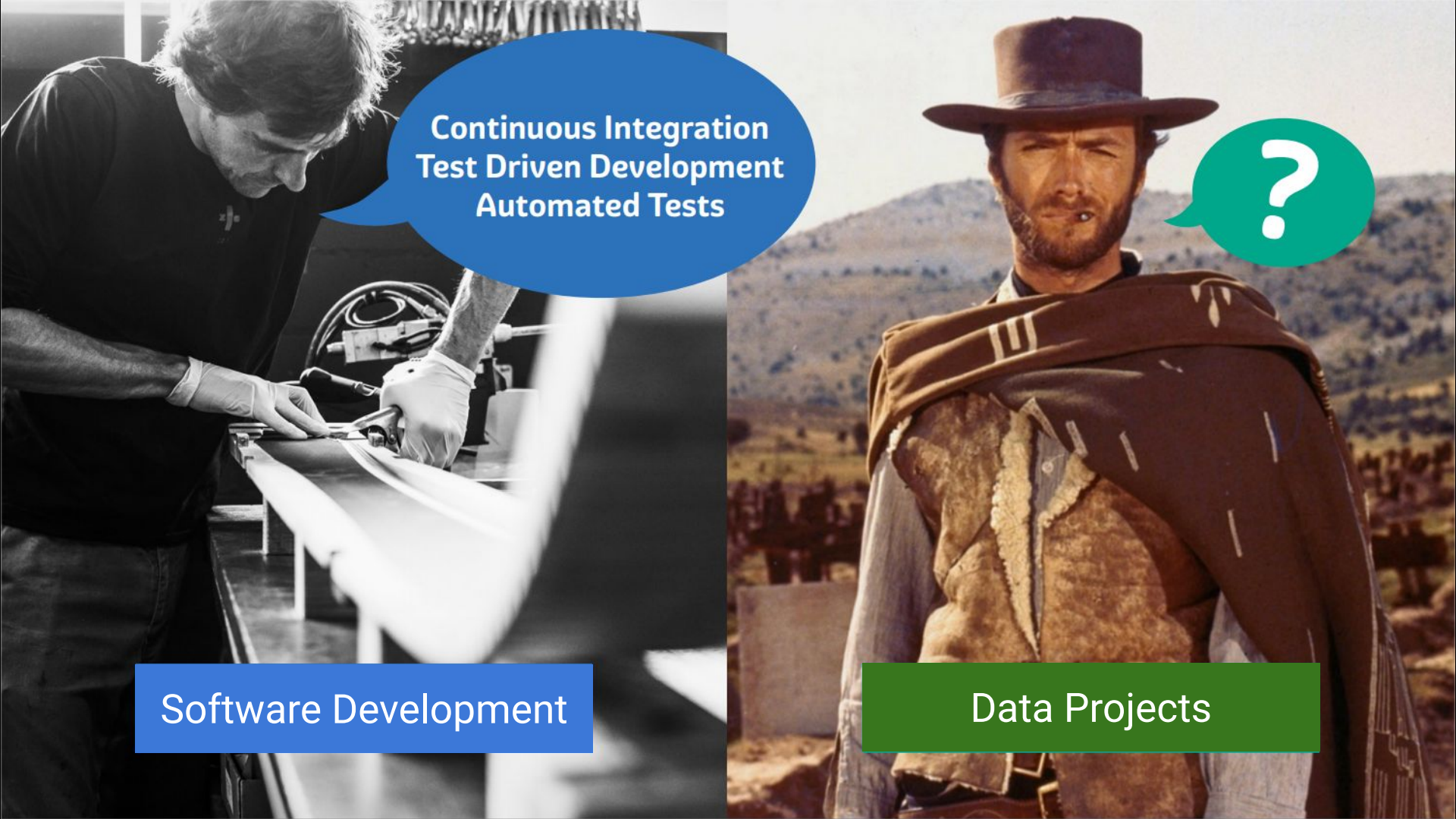
Data Engineering Patterns and Principles

Valdas Maksimavičius


Continuous Integration
Test Driven Development
Automated Tests
Software Development


Microsoft
CODE COMPLETE
2
Second Edition
A practical handbook of software construct
the mythical man-month
ing
The Addison Wesley Signature Series
A KENT BECK SIGNATURE BOOK
TEST-DRIVEN DEVELOPMENT
Design Patterns
Elements of Reusable Object-Oriented Software
Erich Gamma
Richard Helm
ADDISON-WESLEY PROFESSIONAL COMPUTING SERIES
Booch
Architectural Patterns
Uncover essential patterns in the most indispensable realm of enterprise architecture
Domain Driven DESIGN
Tackling Complexity in the Heart of Software


Continuous Integration
Test Driven Development
Automated Tests
?
Software Development
Data Projects

Would you be confident in a self-driving car ...

… knowing that there is your software running it?

Standardize and increase the descriptive power of engineering processes by applying patterns

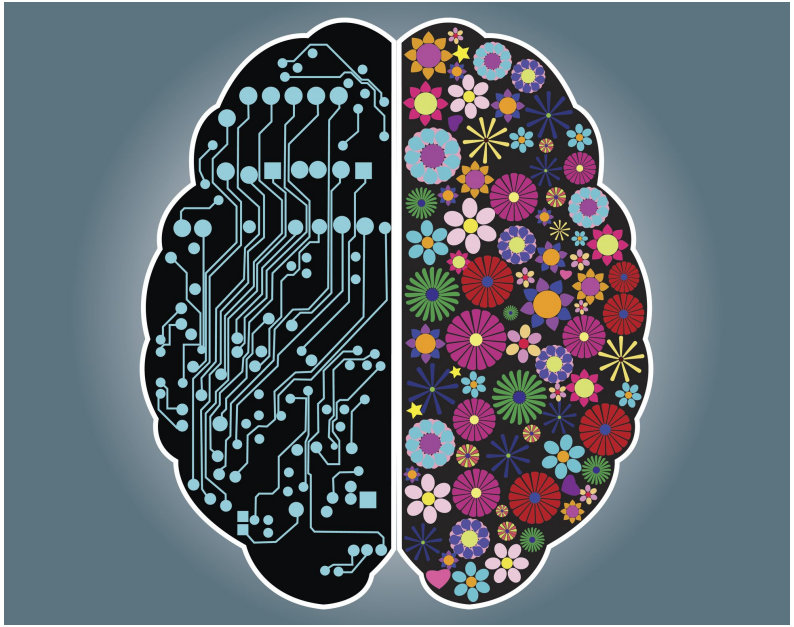
Or in other words

stand on the shoulders of giants

and stop reinventing the wheel

Why does my brain need patterns?

- Left side of your brain is responsible for analytical thinking, science, math, etc.
- It uses known building blocks to model the surrounding world
- **If you like table representation of data, you will try to model everything as a table**
- **As an engineer, expand your tool belt by learning new patterns and new building blocks to solve business problems better.**

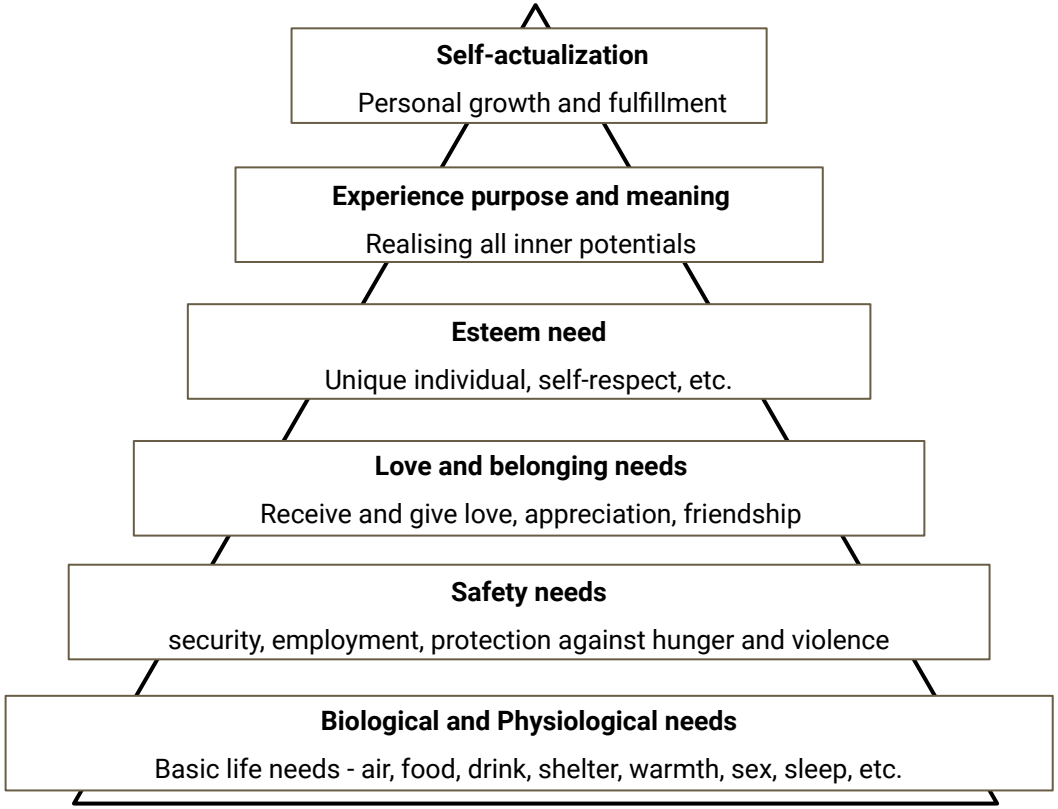


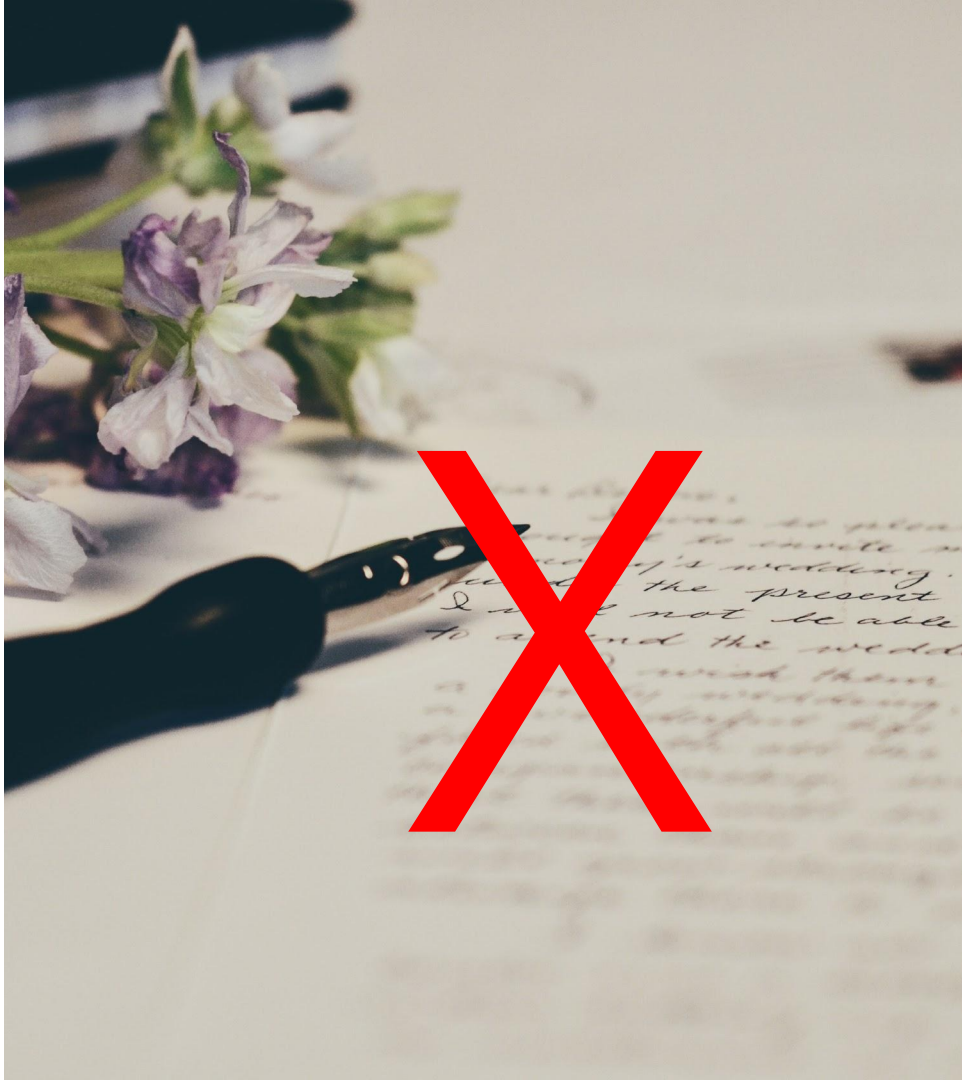
Source: https://www.health.harvard.edu/blog/right-brainleft-brain-right-2017082512222

About me

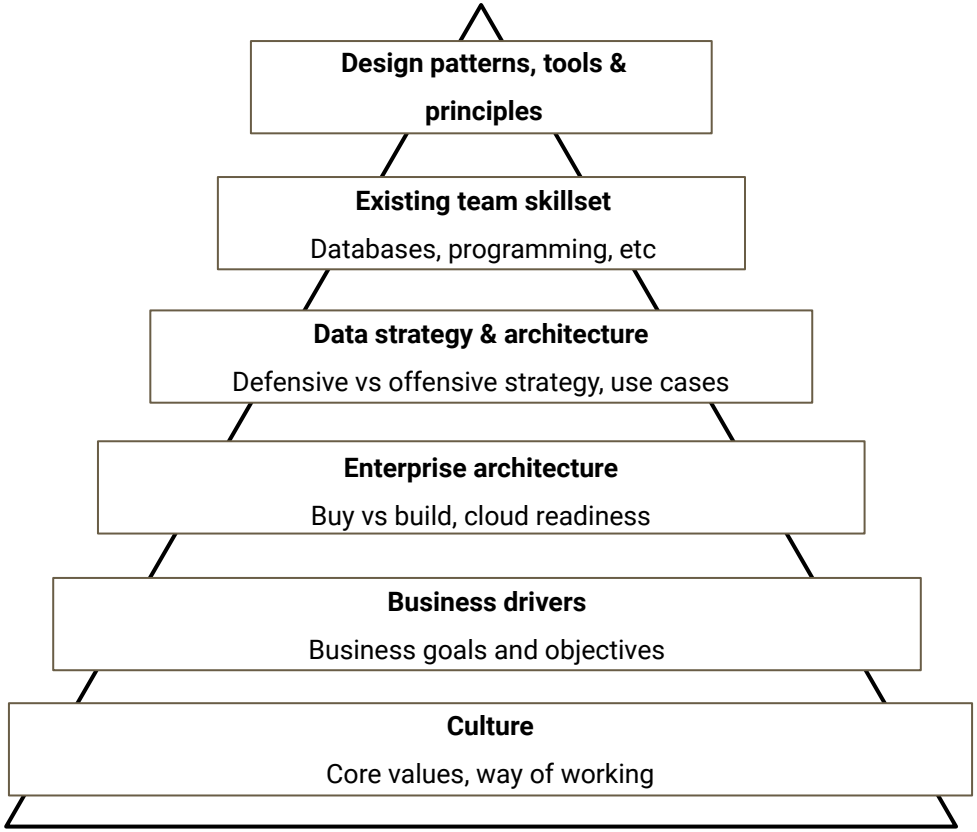
- IT Architect at Cognizant
- Data Engineering, Data Science, Cloud Computing, Agile teams
- Financial, Manufacturing, Logistics, Retail industries
- Organizer of Vilnius Microsoft Data Platform Meetup & Hack4Vilnius Hackathon
- Blogging on *www.valdas.blog*

Maslow’s hierarchy of needs

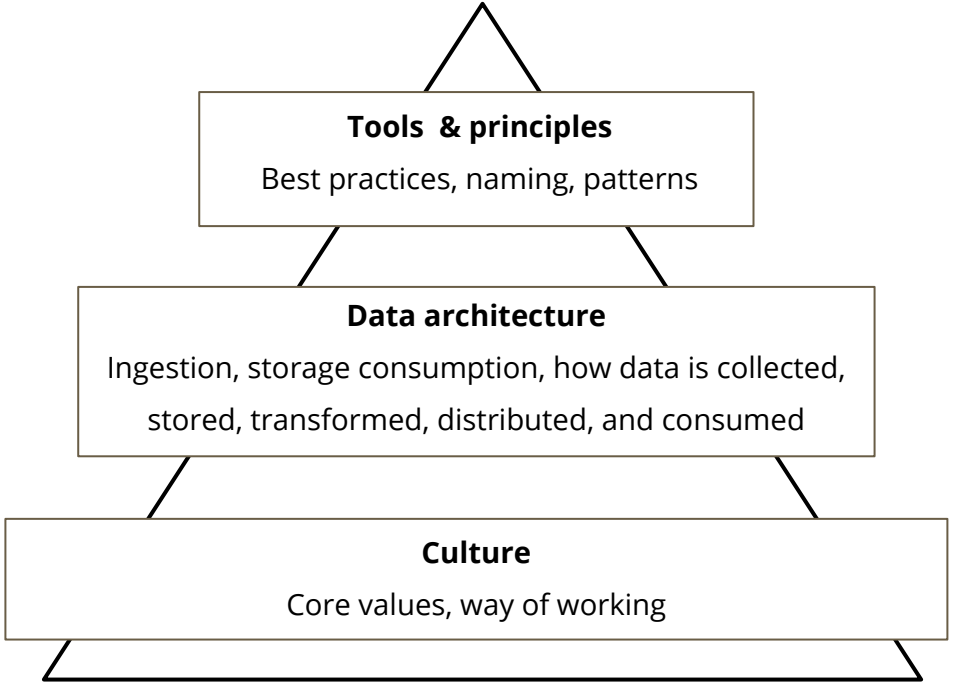




Maslow's hierarchy of needs for data projects



Maslow's hierarchy of needs for data projects - simplified view for today's presentation

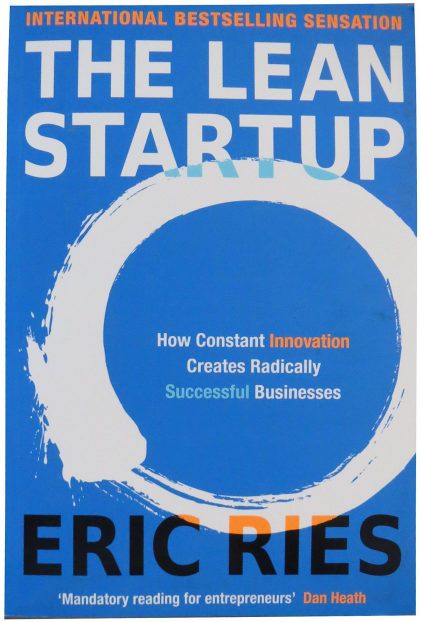


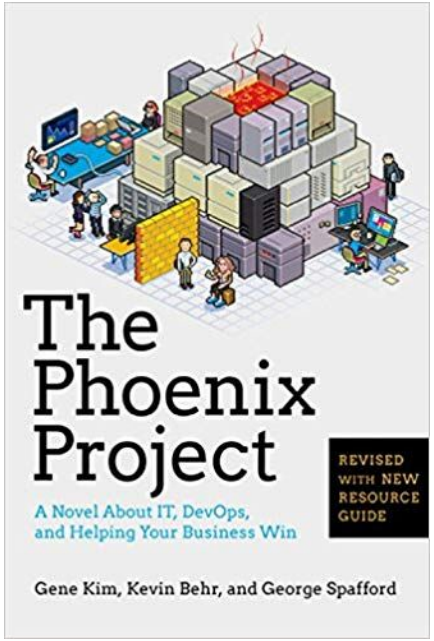
Culture, way of working, values

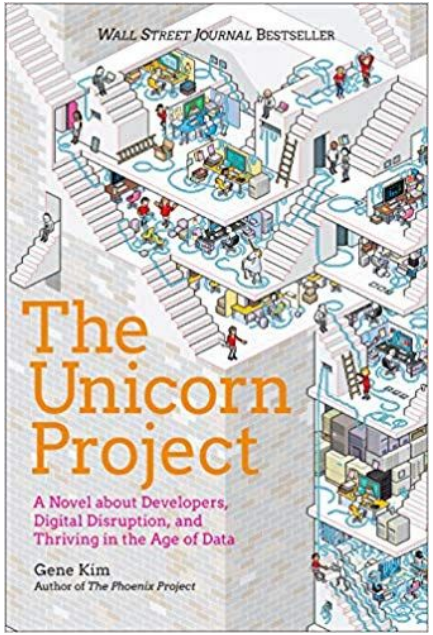
DevOps culture

1. Foster a Collaborative Environment
2. Impose End-to-End Responsibility - you build it you ship it
3. Encourage Continuous Improvement
4. Automate (Almost) Everything
5. Focus on the Customer’s Needs
6. Embrace Failure, and Learn From it
7. Unite Teams — and Expertise

Source: https://www.cmswire.com/information-management/7-key-principles-for-a-successful-devops-culture/


INTERNATIONAL BESTSELLING SENSATION
THE LEAN STARTUP
How Constant Innovation Creates Radically Successful Businesses
ERIC RIES
'Mandatory reading for entrepreneurs' Dan Heath


The Phoenix Project
REVISED WITH NEW RESOURCE GUIDE
A Novel About IT, DevOps, and Helping Your Business Win
Gene Kim, Kevin Behr, and George Spafford

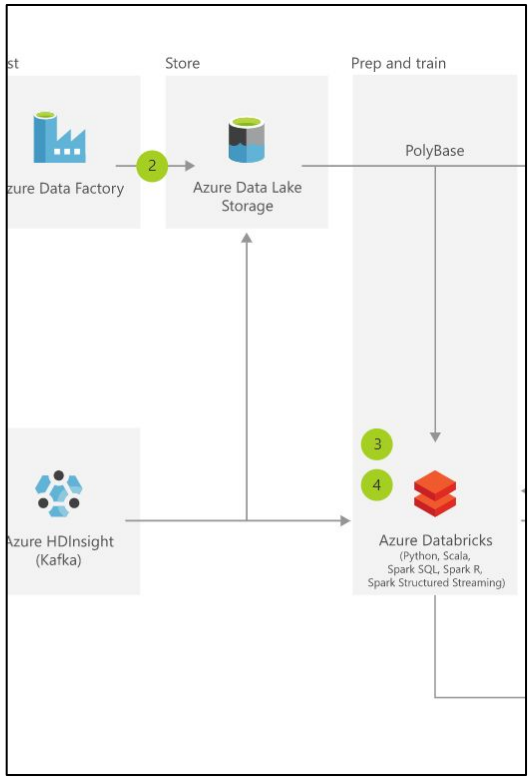

WALL STREET JOURNAL BESTSELLER
The Unicorn Project
A Novel about Developers, Digital Disruption, and Thriving in the Age of Data
Gene Kim
Author of The Phoenix Project

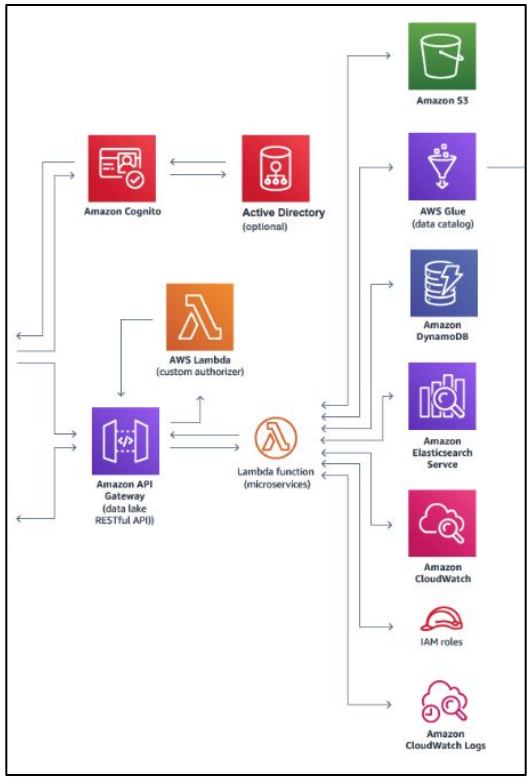
Data architecture

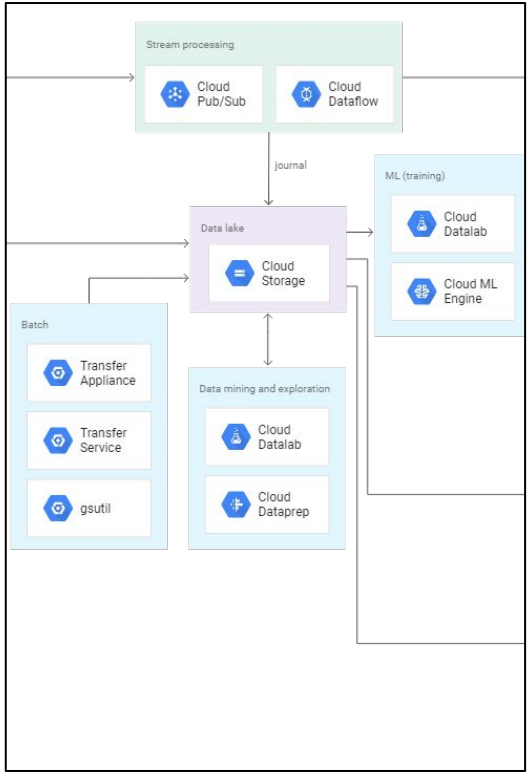
If you are building a data platform in the cloud, remember that ...

low barrier-to-entry overshadows complexity

Big Data cloud architecture references

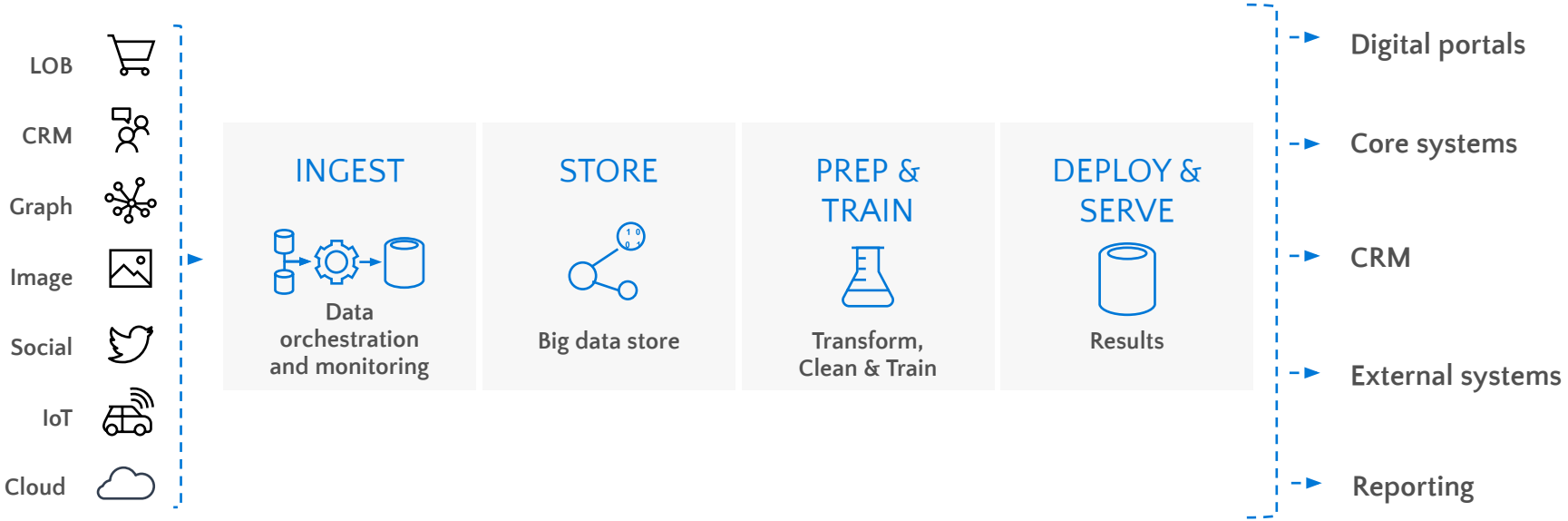




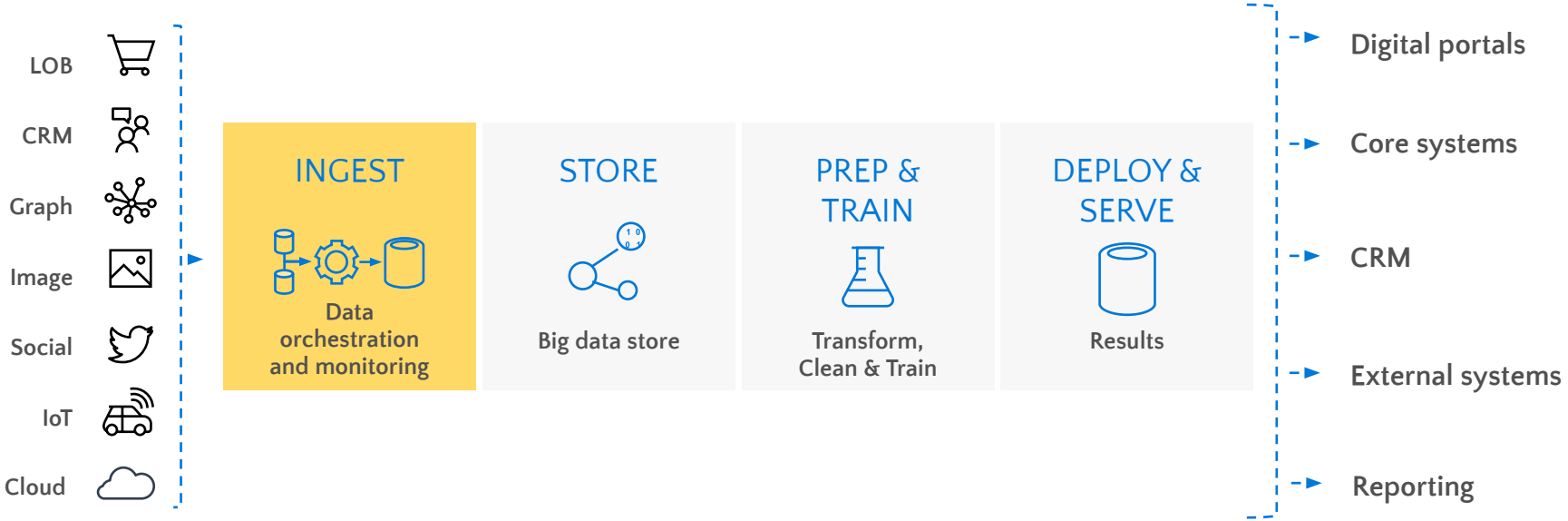


Source: https://azure.microsoft.com/en-in/solutions/architecture/modern-data-warehouse/

Architecture example



Data ingestion



Application integration approaches

File Transfer

Have each application produce files of shared data for others to consume, and consume files that others have produced.

Shared Database

Have the applications store the data they wish to share in a common database.

Remote Procedure Invocation

Have each application expose some of its procedures so that they can be invoked remotely, and have applications invoke those to run behavior and exchange data.

Messaging

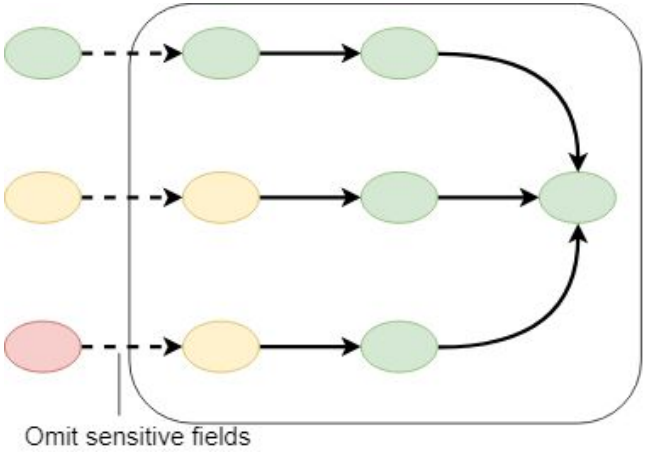
Have each application connect to a common messaging system, and exchange data and invoke behavior using messages.

Ingestion challenges

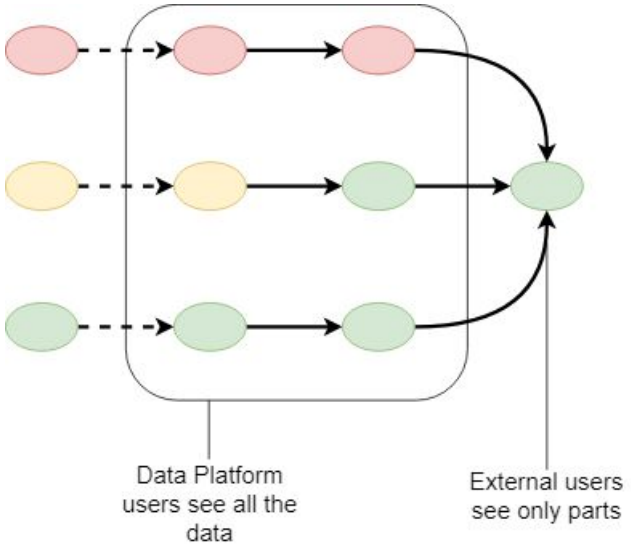
- Multiple data source load and prioritization -> **push vs pull strategy**
- Ingested data indexing and tagging -> **metadata collection is mandatory**
- Data validation and cleansing -> **separate business from processing logic**
- Data transformation and compression -> **different compression and file types**

Choose privacy protection patterns

Privacy protection at the ingress

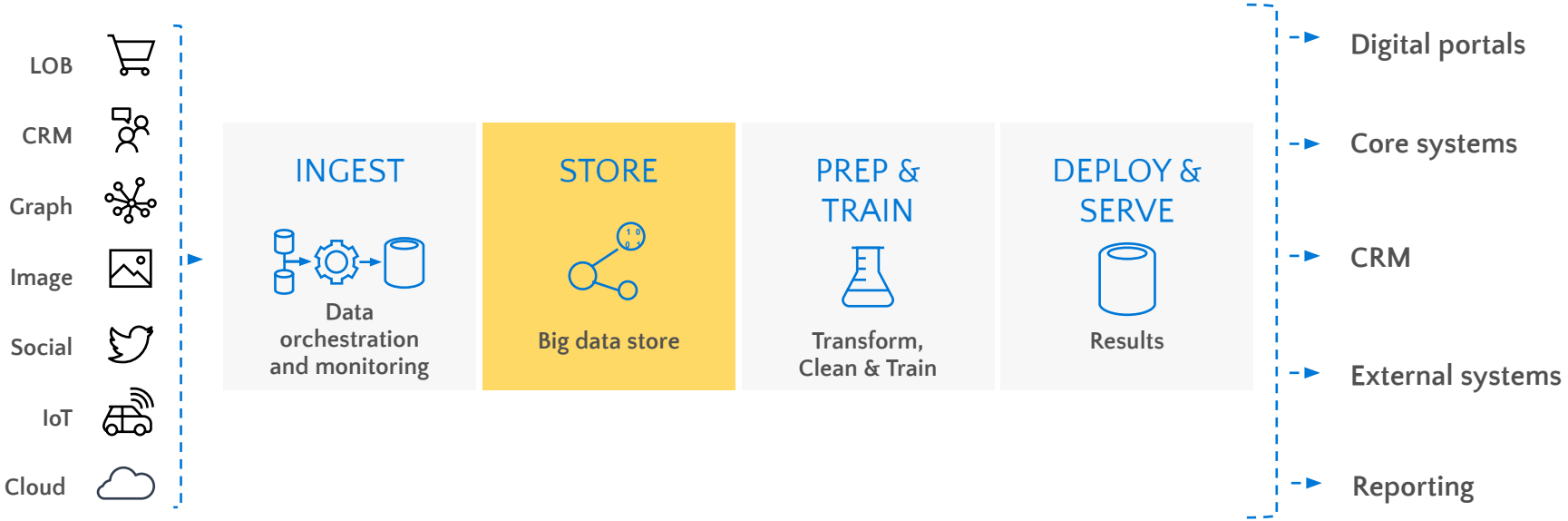


Privacy protection at the egress



Source: https://www.valdas.blog/2019/08/06/privacy-gdpr-implementation-in-azure/

Data storage



Use cloud storage offerings instead of Hadoop

Machine Learning & Big Data Blog

Is Hadoop Dead? How Kubernetes and Cloud-Native Could Displace Hadoop

The Death of Hadoop?

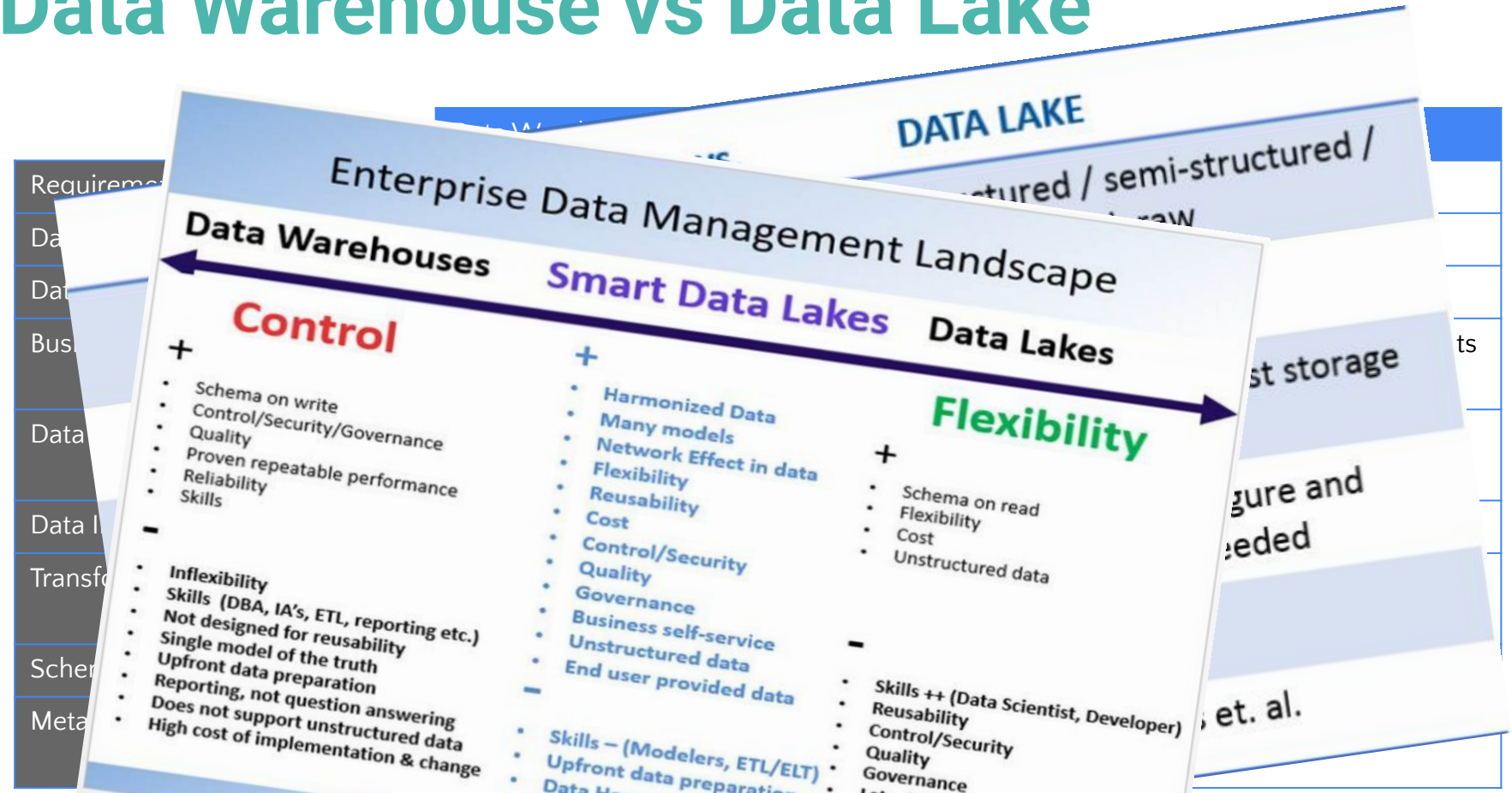
Is Hadoop dead? Not so fast. Plan on supporting multiple environments for some time to come.

By Barry Devlin

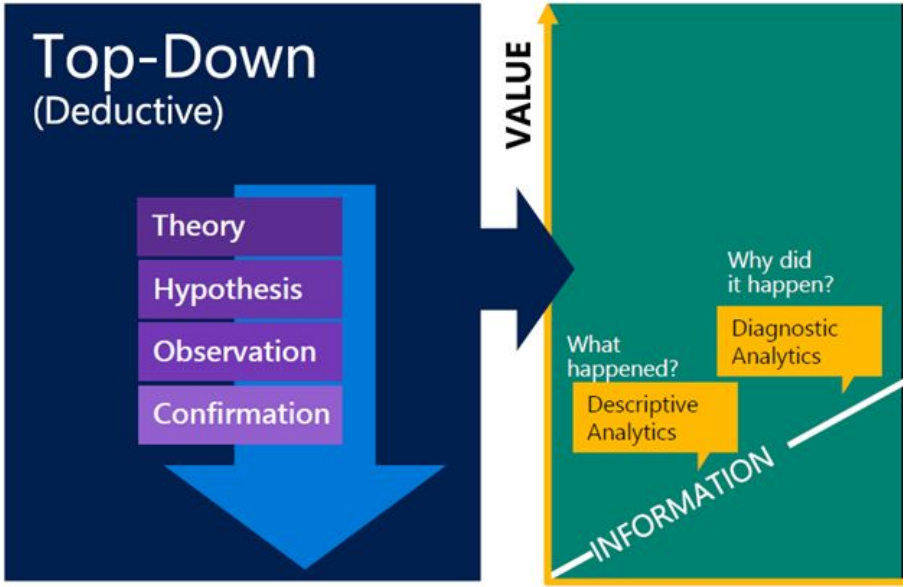
February 26, 2019

The recent "merger of equals" between Cloudera and Hortonworks has triggered speculation about the possible imminent demise of Hadoop. Market observers question if the merger indicates a shrinking Hadoop ecosystem market that can no longer support its two largest competing beasts.

Data Warehouse vs Data Lake

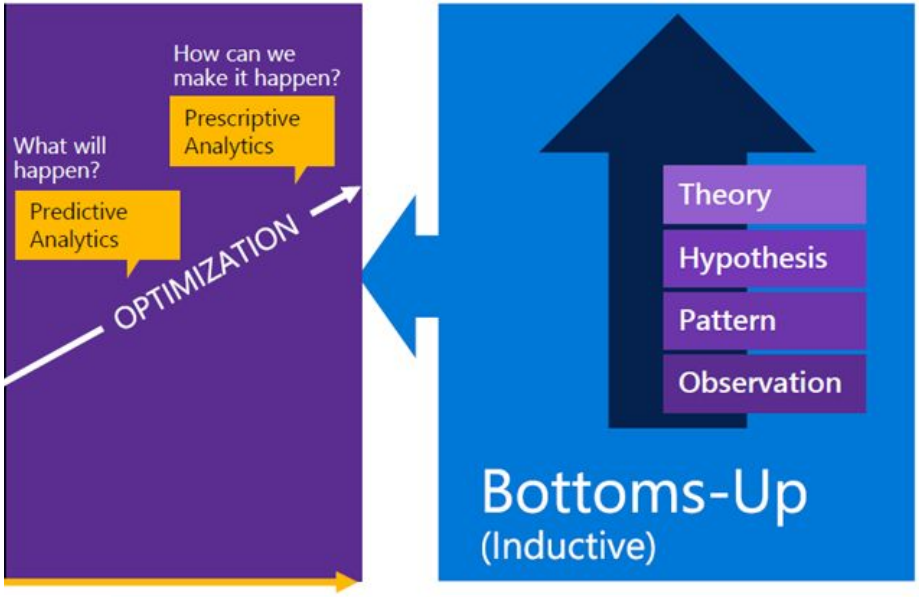


Data Warehouse vs Data Lake



- Know the questions to ask
- Lot's of upfront work to get the data to where you can use it
- Model first

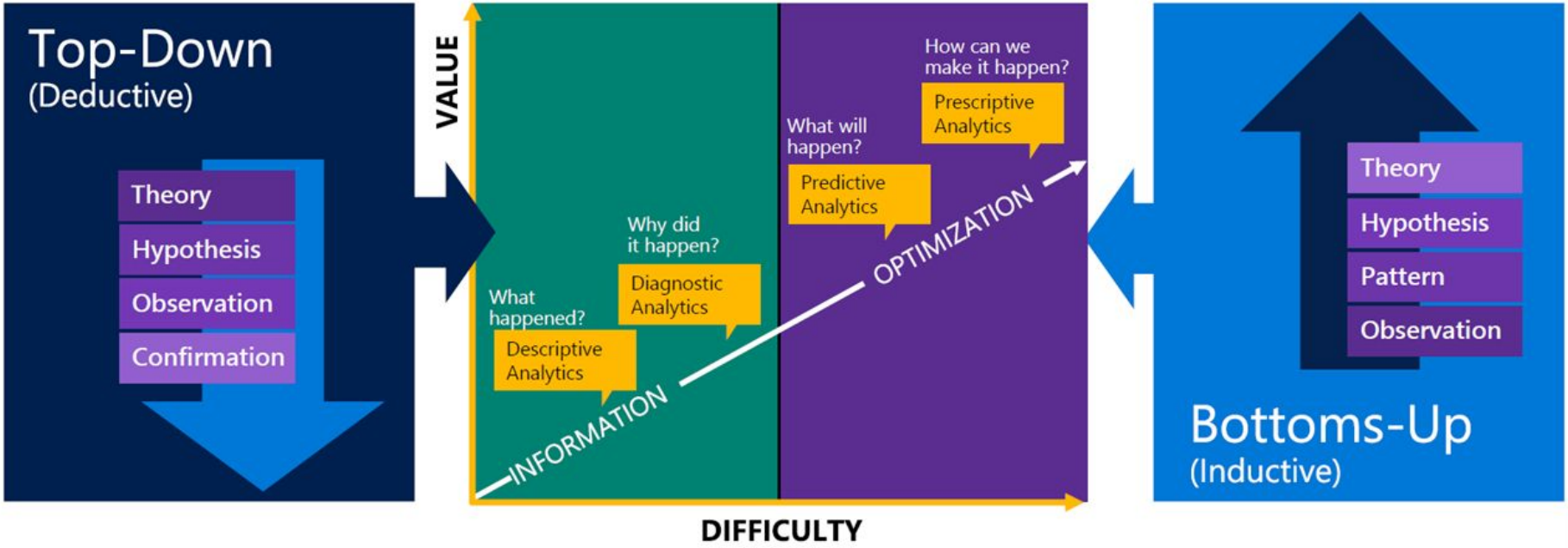
Data Warehouse vs Data Lake



- Don't know the questions to ask
- Little upfront work needs to be done to start using data
- Model later

Source: Microsoft

Data Warehouse vs Data Lake

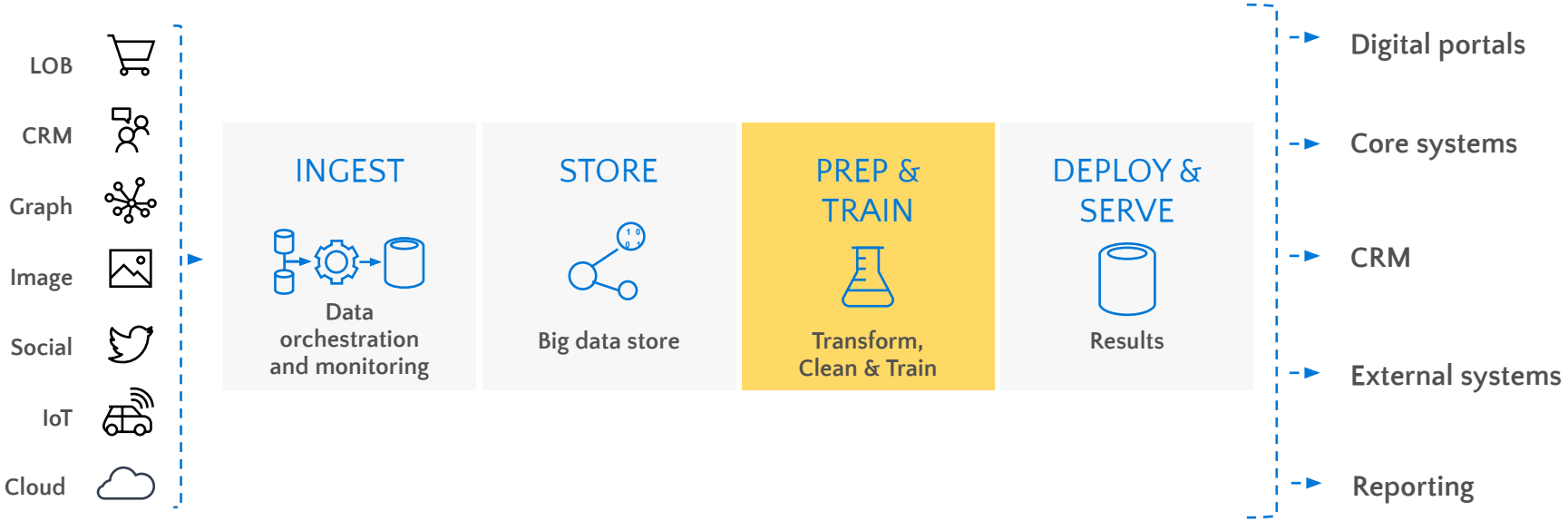


- Know the questions to ask
- Lot's of upfront work to get the data to where you can use it
- Model first

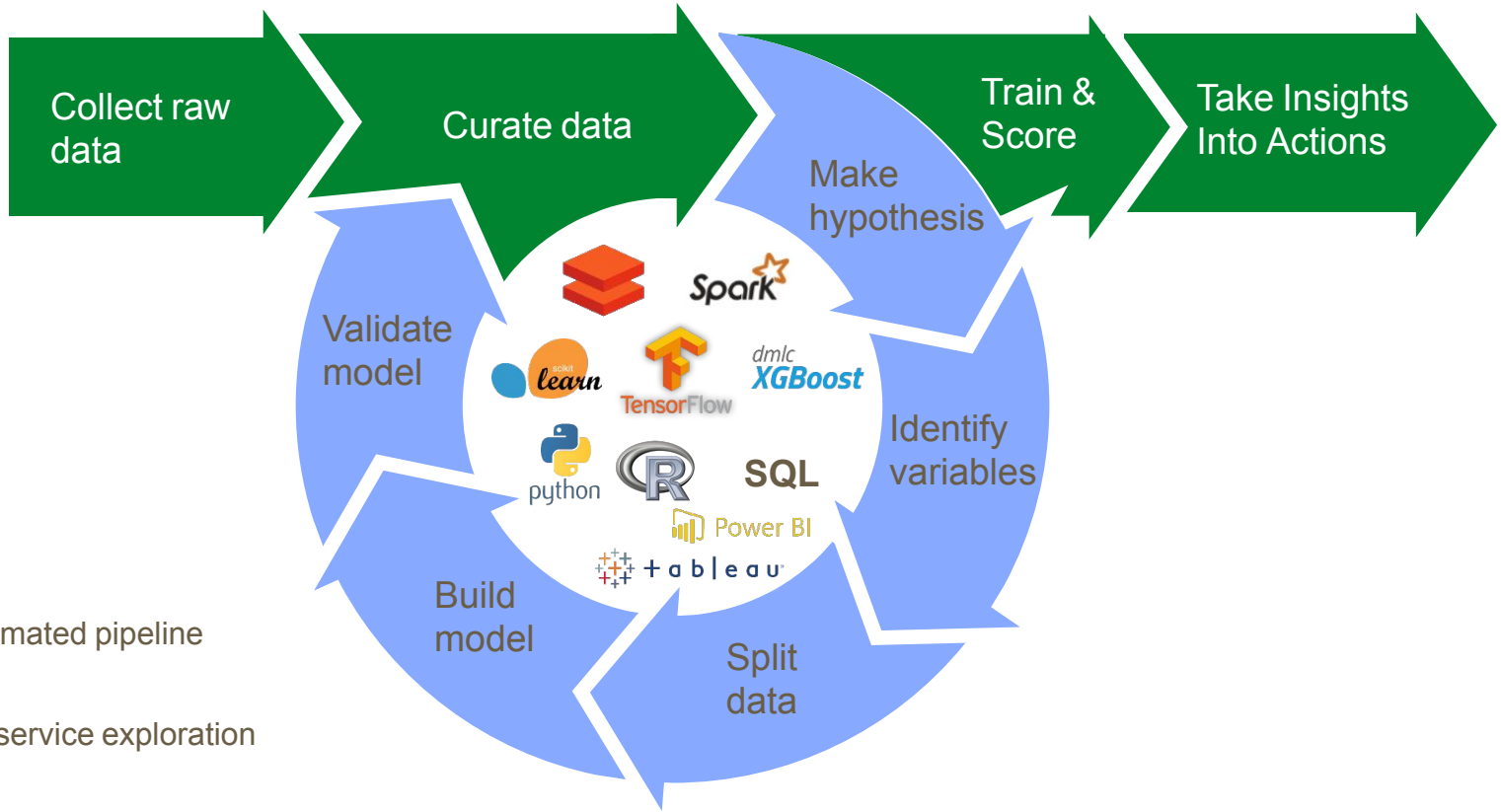
- Don't know the questions to ask
- Little upfront work needs to be done to start using data
- Model later

Source: Microsoft

Data preparation & training



Offer self-service tools



Use on-demand resources

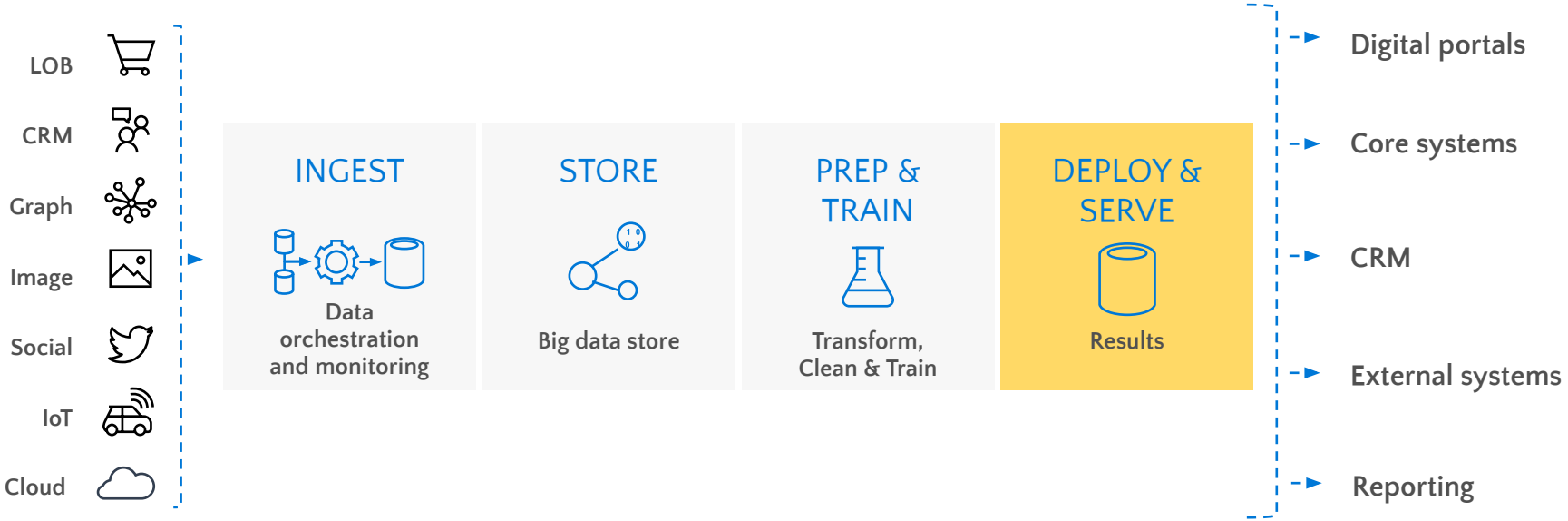
Storage

10 TB
+ 24/7
= ~300 Eur/month

databricks

32 cores / 128 GB RAM
+ 160 hours
= ~700 Eur/month

Serve results to end consumers



Apply domain and product thinking

- Model to describe a domain
- Unified language
- Raw or transformed datasets
- Domain team is responsible for its lifecycle, SLA
- Discoverable, addressable, trustworthy, self-describing, interoperable, secure
- Each producer is responsible of sharing data products to organization

How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh

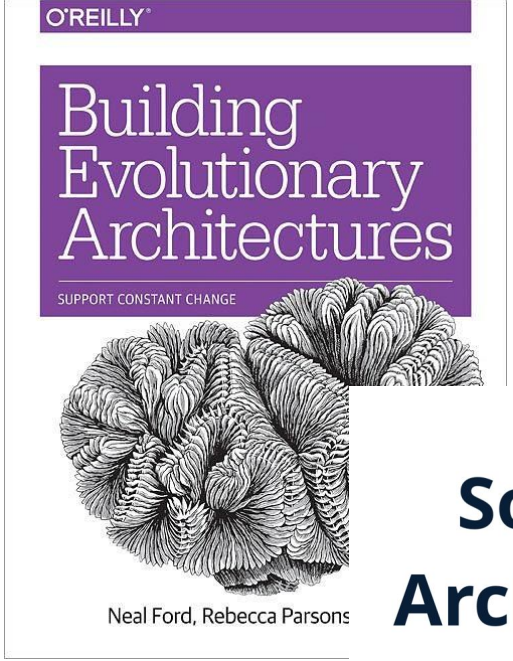
Many enterprises are investing in their next generation data lake, with the hope of democratizing data at scale to provide business insights and ultimately make automated intelligent decisions. Data platforms based on the data lake architecture have common failure modes that lead to unfulfilled promises at scale. To address these failure modes we need to shift from the centralized paradigm of a lake, or its predecessor data warehouse. We need to shift to a paradigm that draws from modern distributed architecture: considering domains as the first class concern, applying platform thinking to create self-serve data infrastructure, and treating data as a product.

20 May 2019

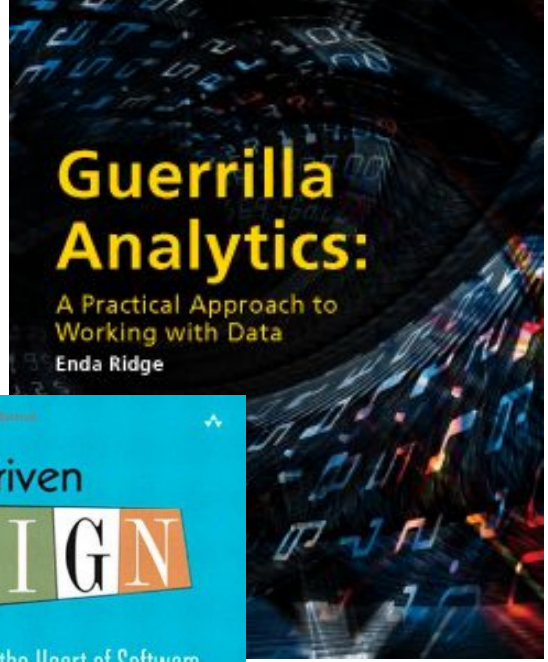
Zhamak Dehghani

Zhamak is a principal technology consultant at ThoughtWorks with a focus on distributed systems architecture and digital platform strategy at Enterprise. She is a member of ThoughtWorks Technology Advisory Board and contributes to the

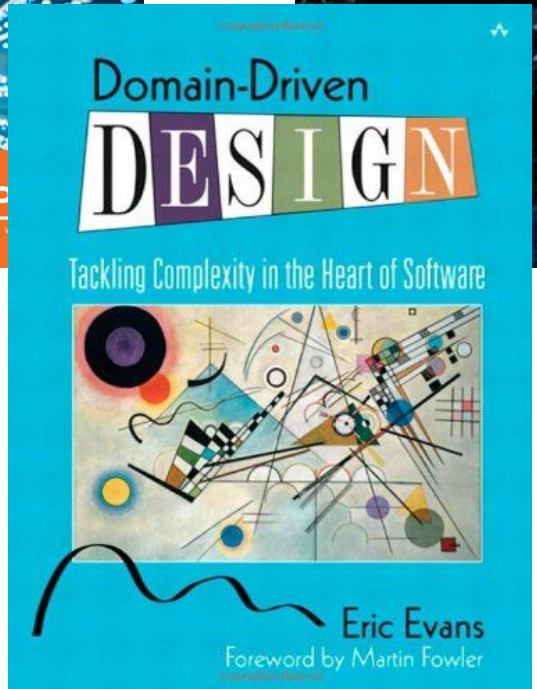
CONTENTS

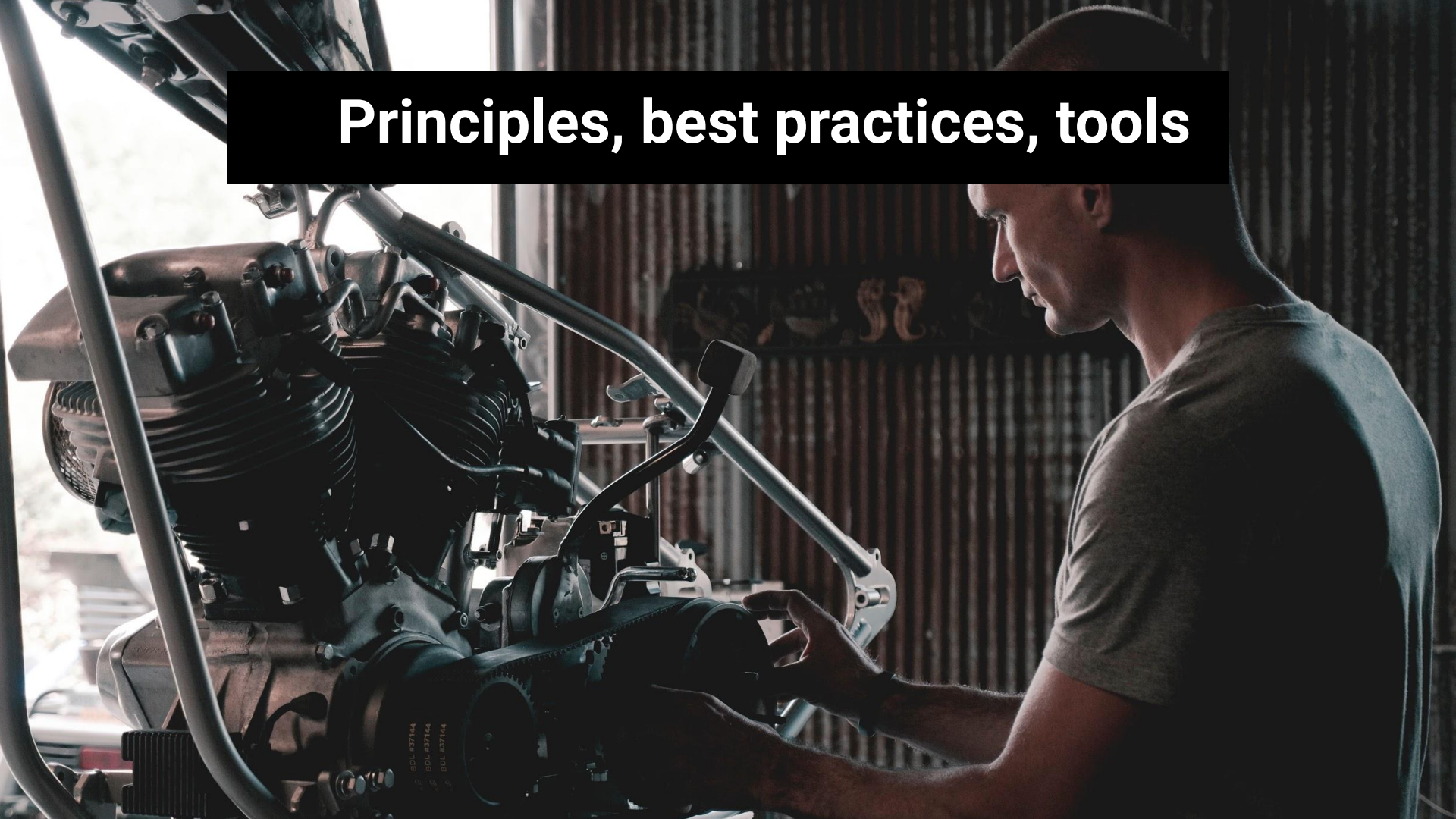

O'REILLY
Building Evolutionary Architectures
SUPPORT CONSTANT CHANGE
Neal Ford, Rebecca Parsons


Architectural Patterns
Uncover essential patterns in the most indispensable realm of enterprise architecture

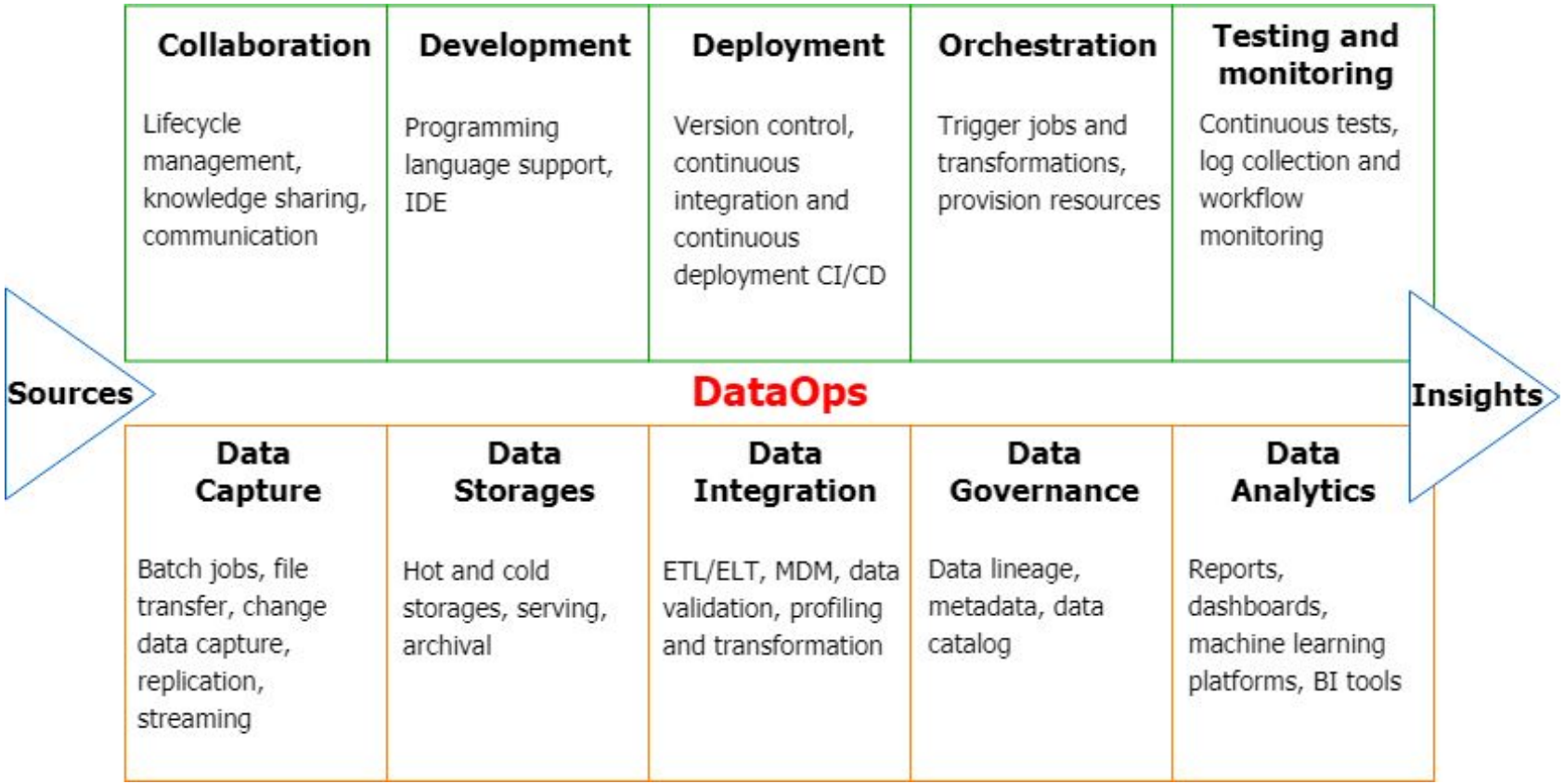

Guerrilla Analytics:
A Practical Approach to Working with Data
Enda Ridge


Software Architecture
for Developers
Volume I
Technical leadership and the balance with agility
Simon Brown

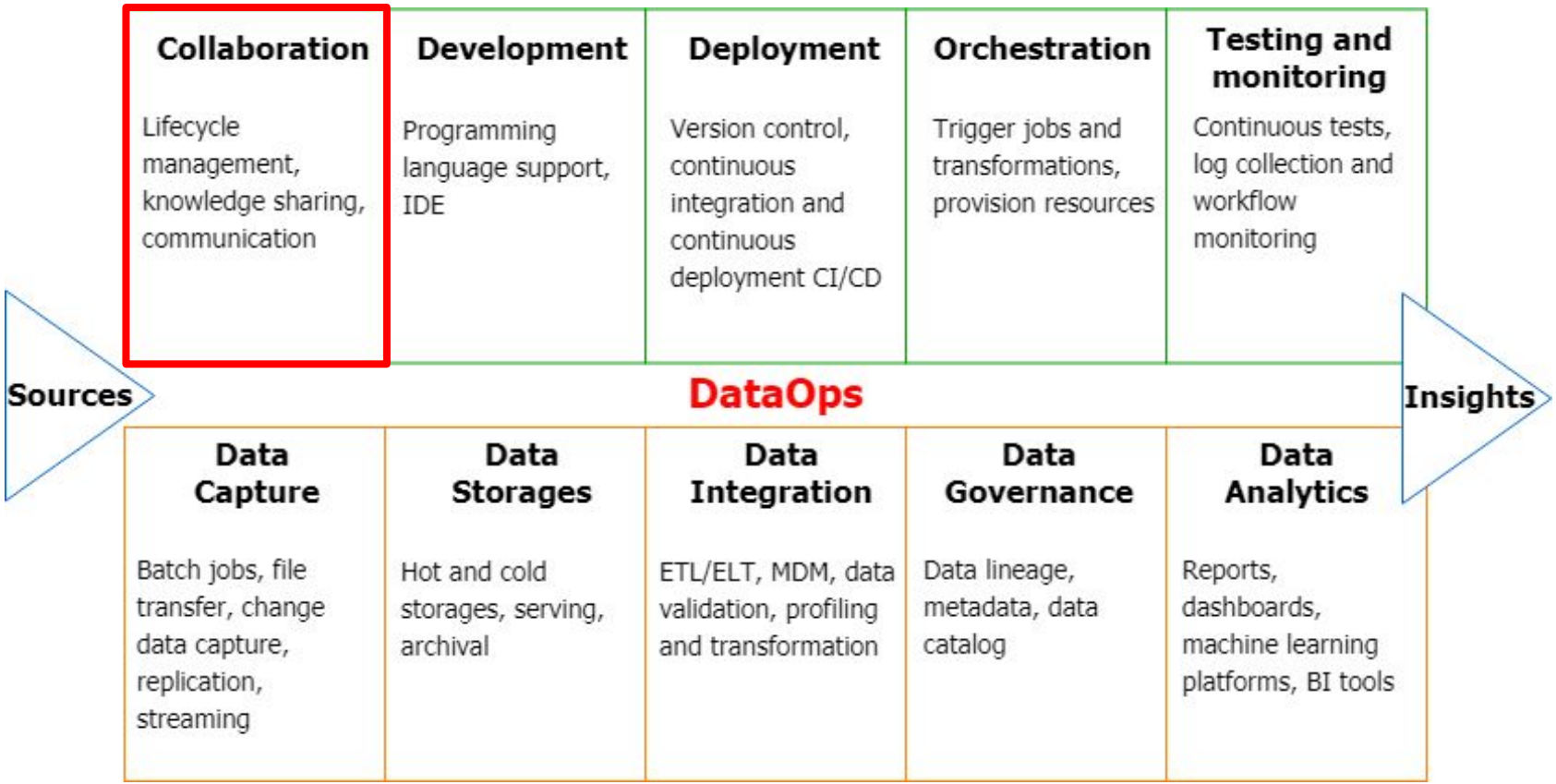

Domain-Driven DESIGN
Tackling Complexity in the Heart of Software
Eric Evans
Foreword by Martin Fowler


Principles, best practices, tools

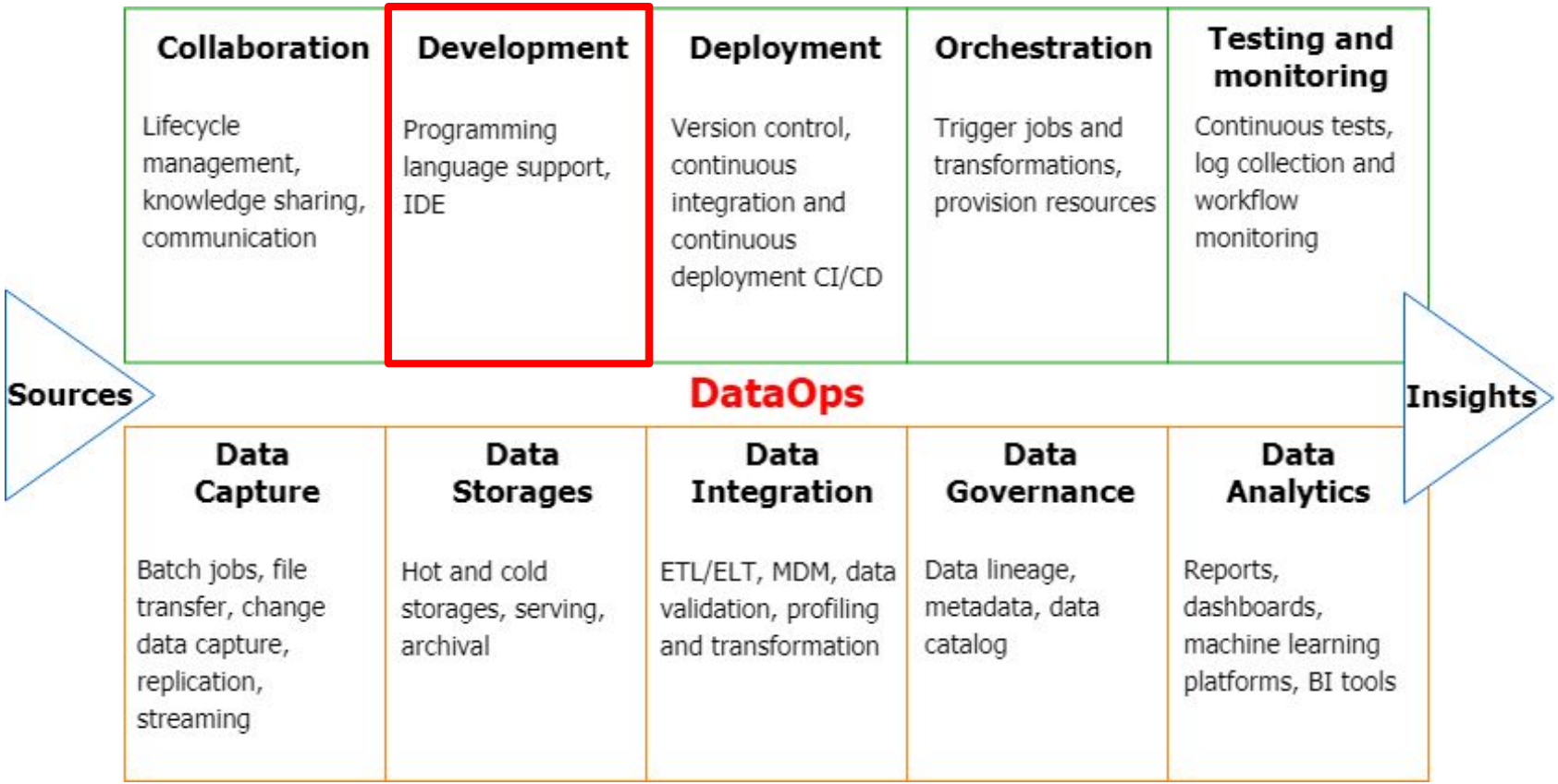
Get familiar with DataOps



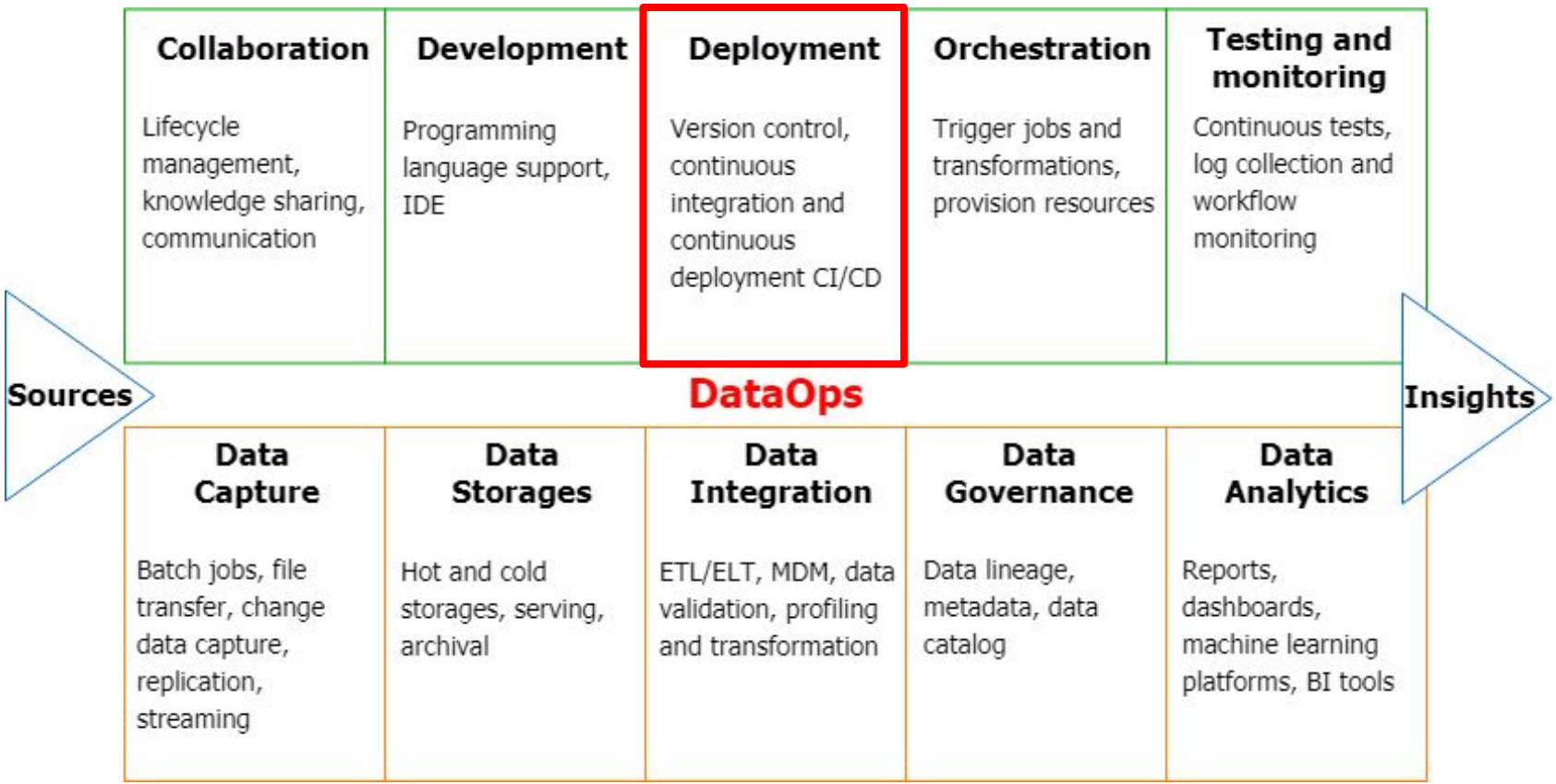
Get familiar with DataOps



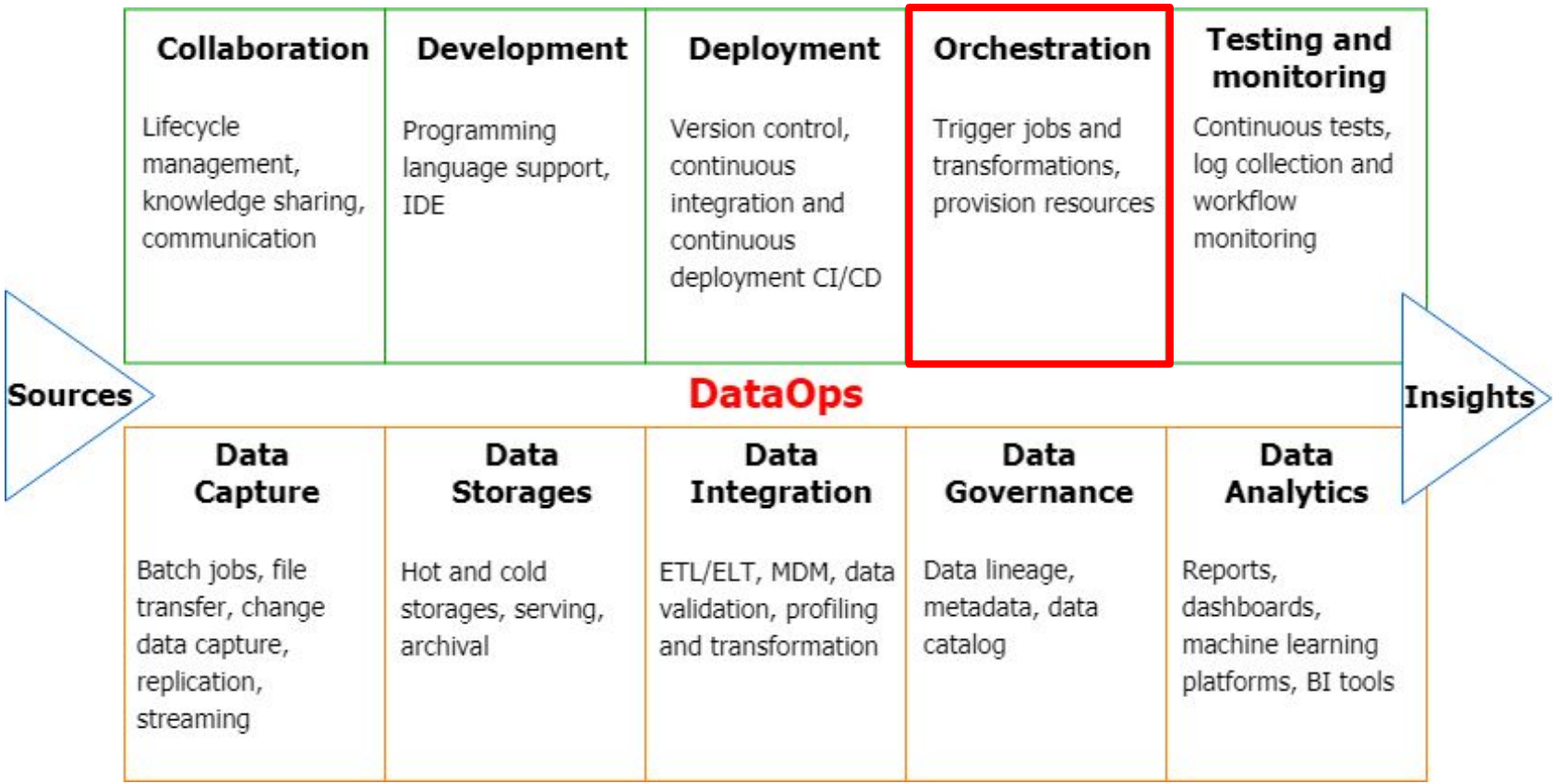
Get familiar with DataOps



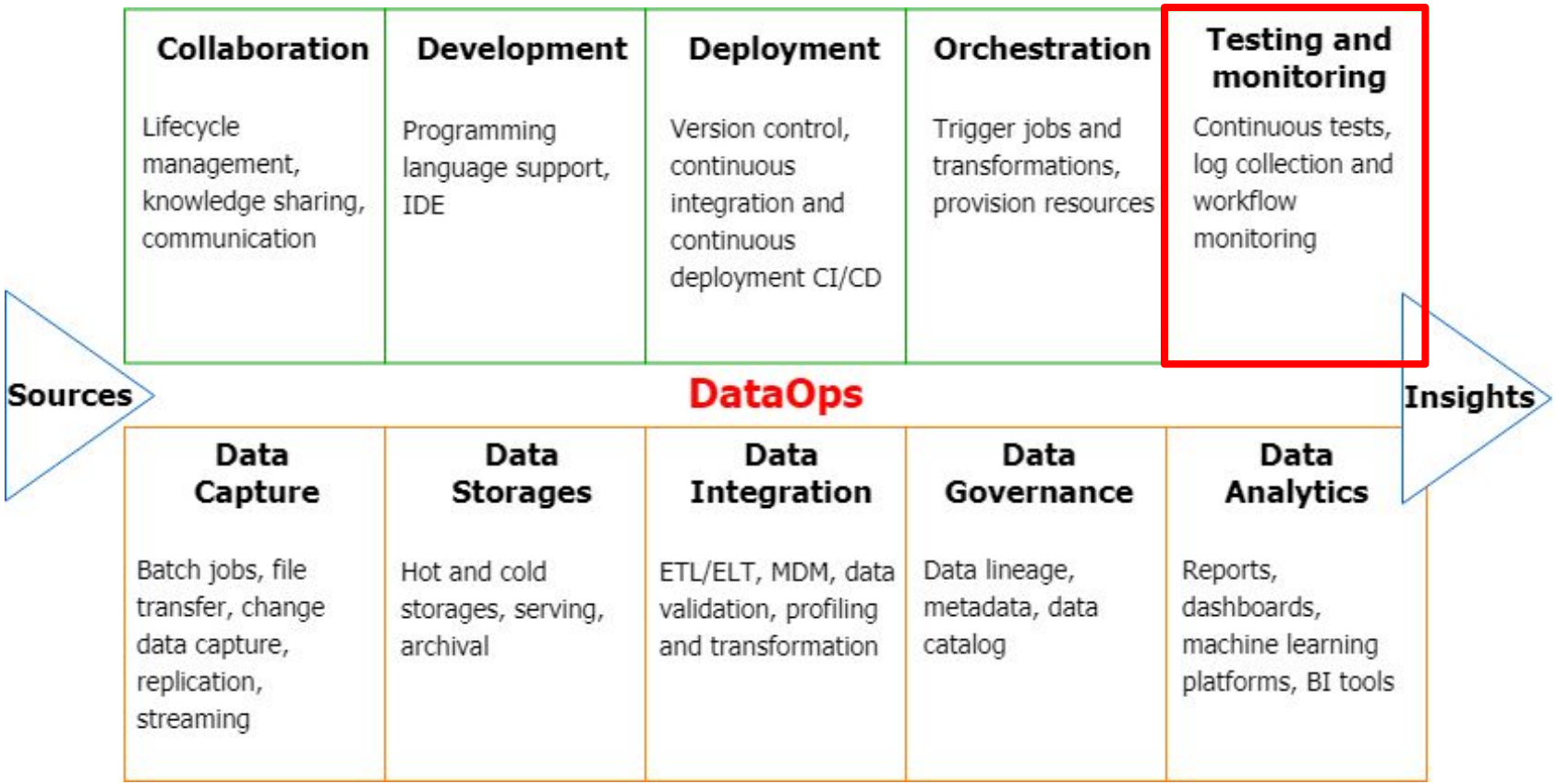
Get familiar with DataOps



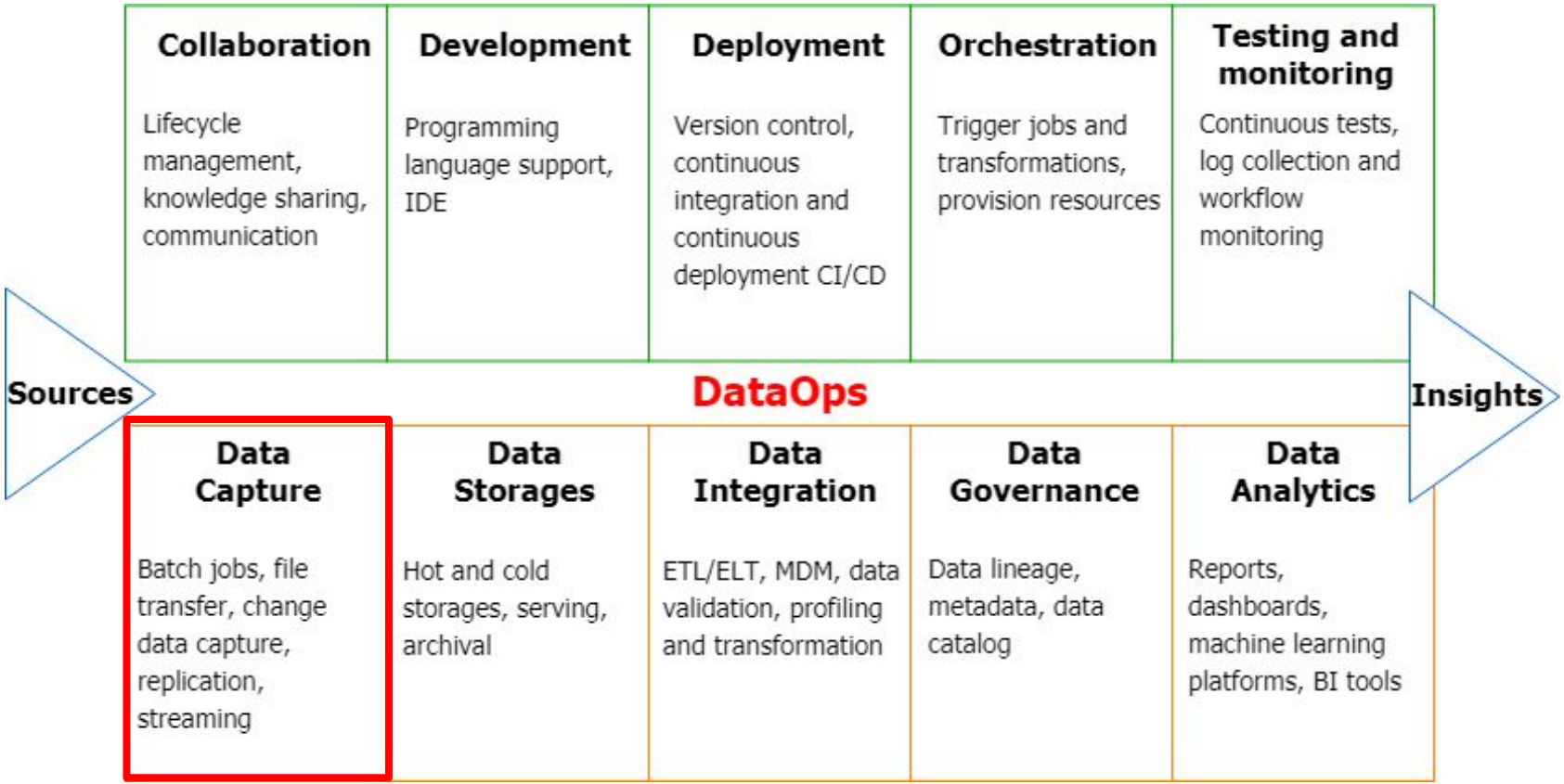
Get familiar with DataOps



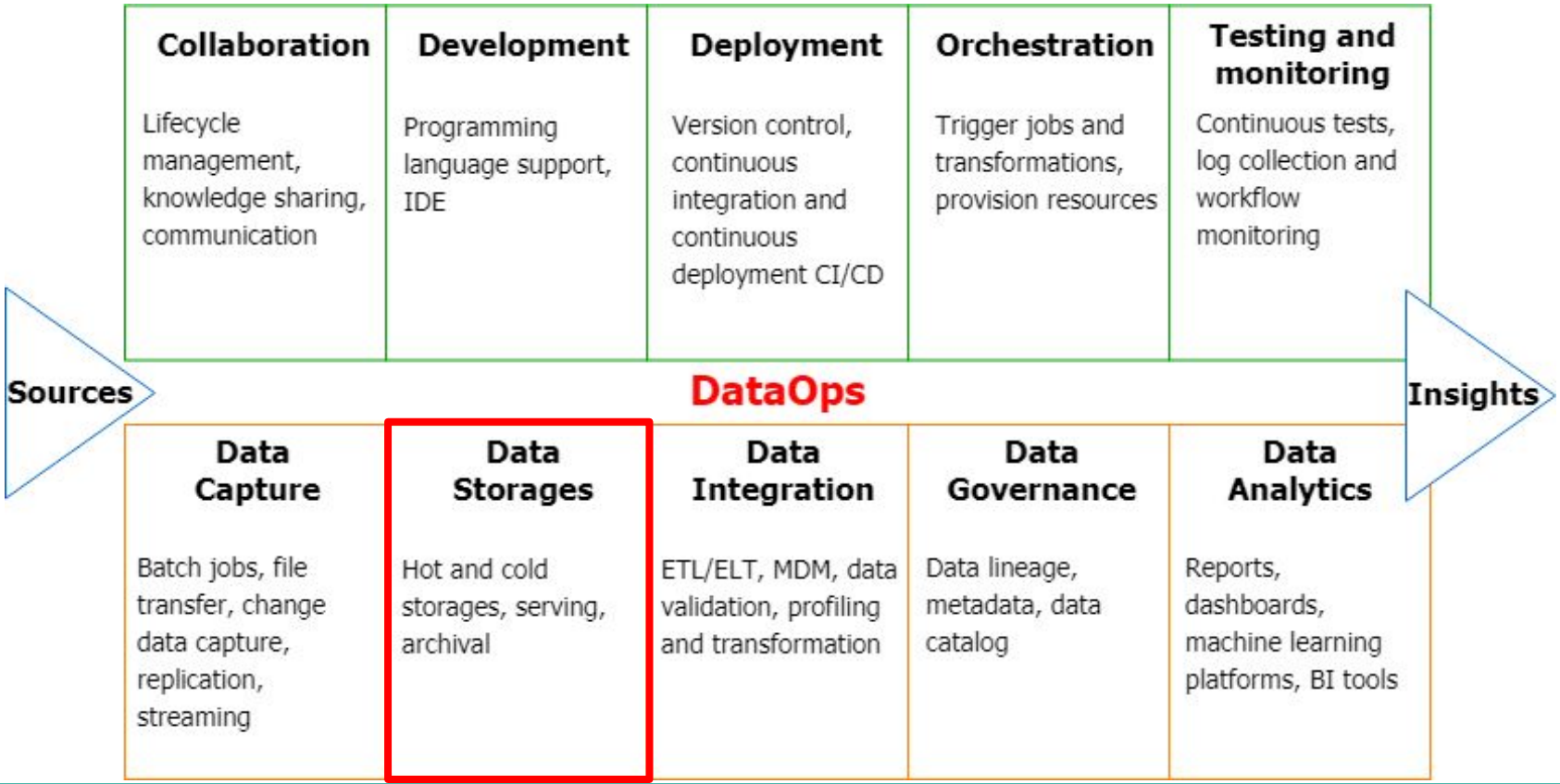
Get familiar with DataOps



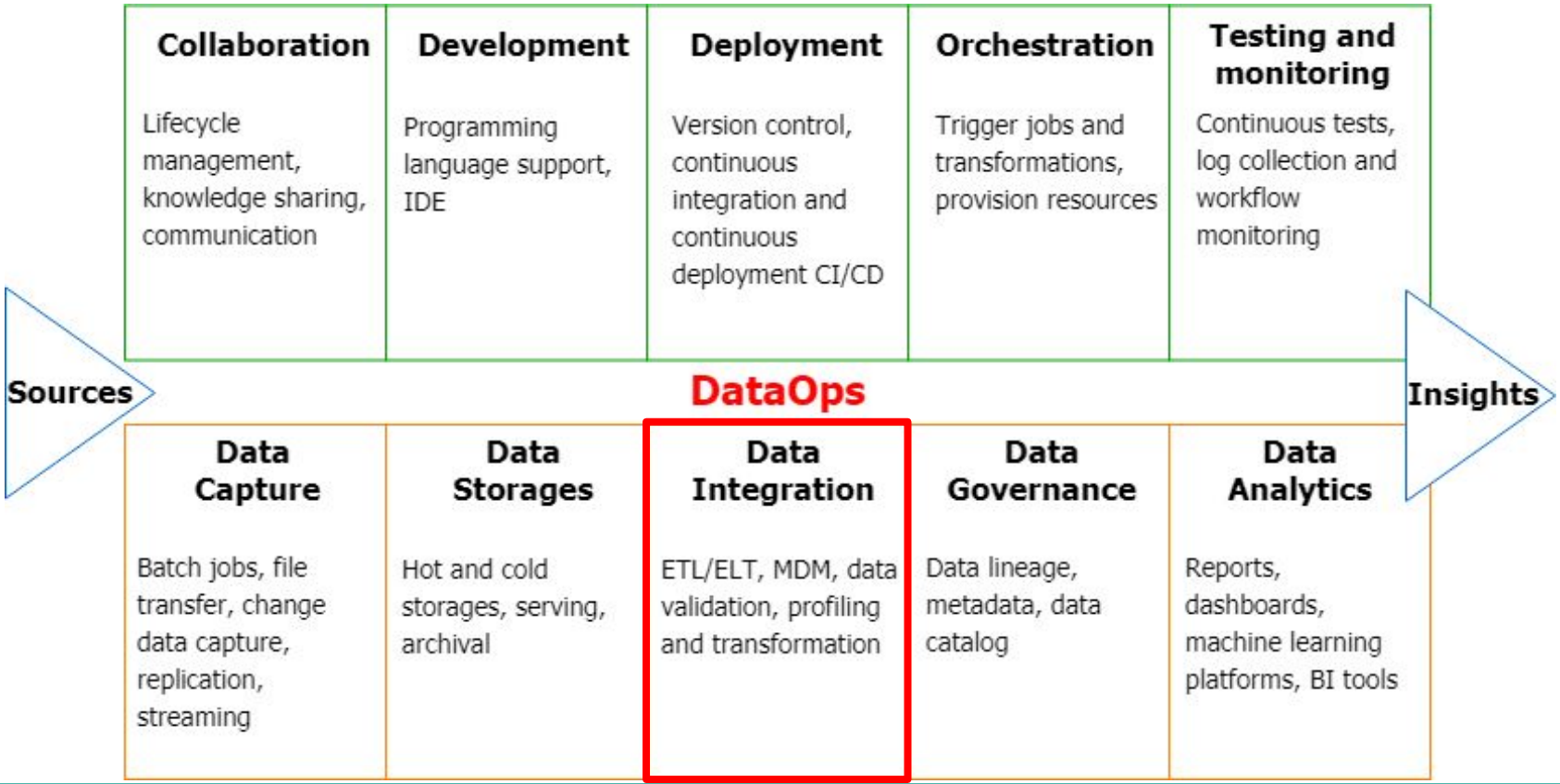
Get familiar with DataOps



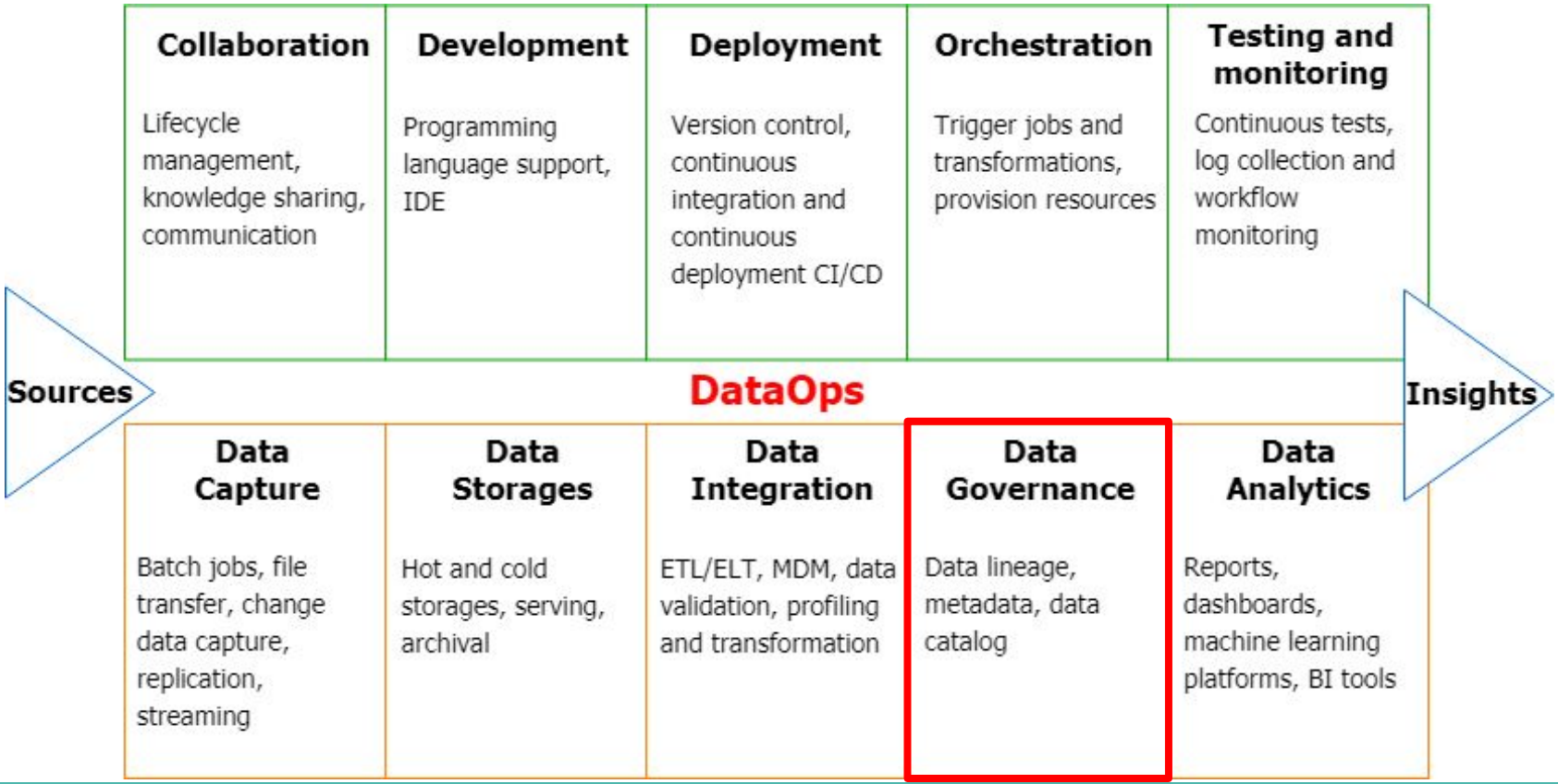
Get familiar with DataOps



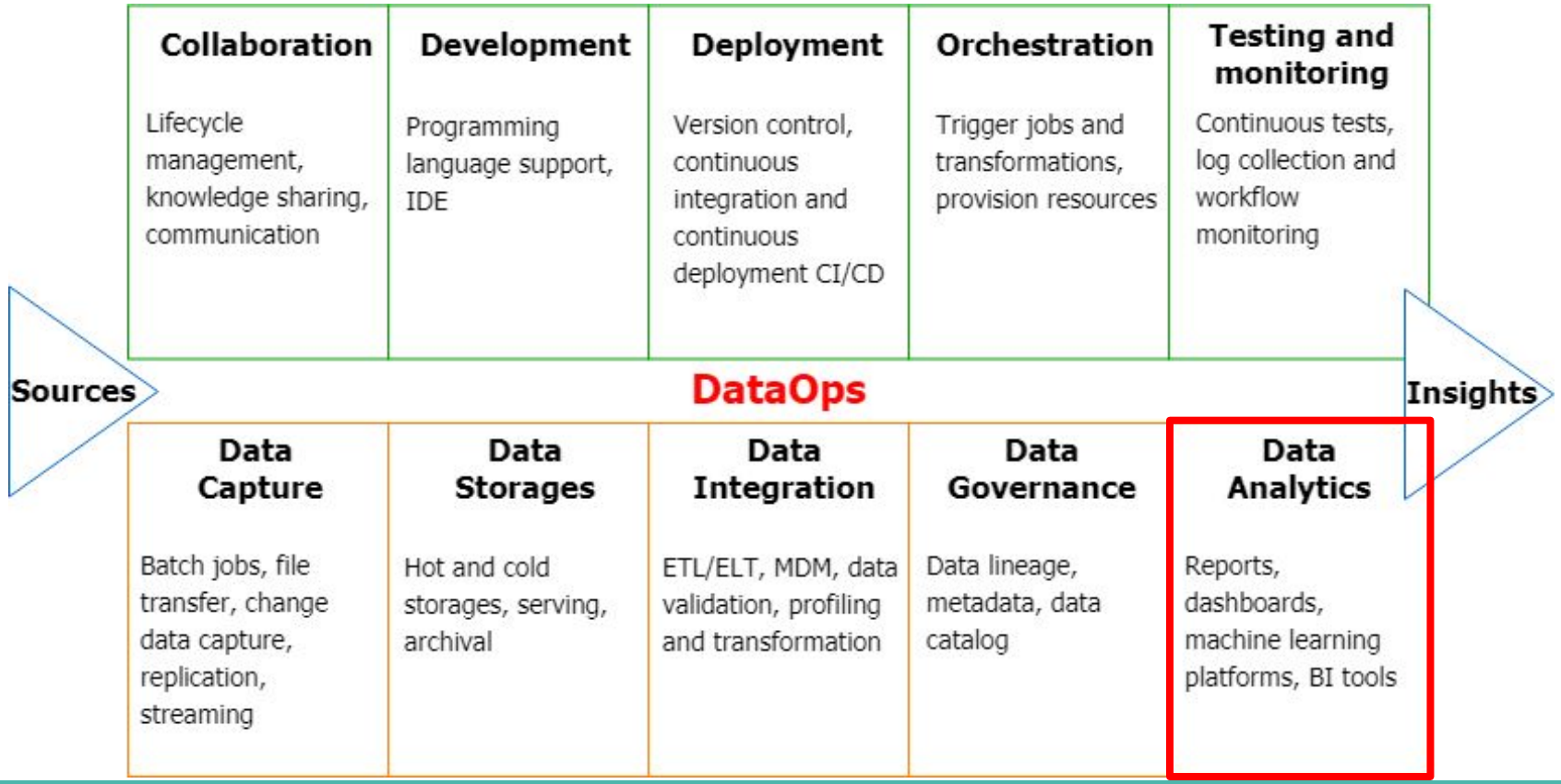
Get familiar with DataOps



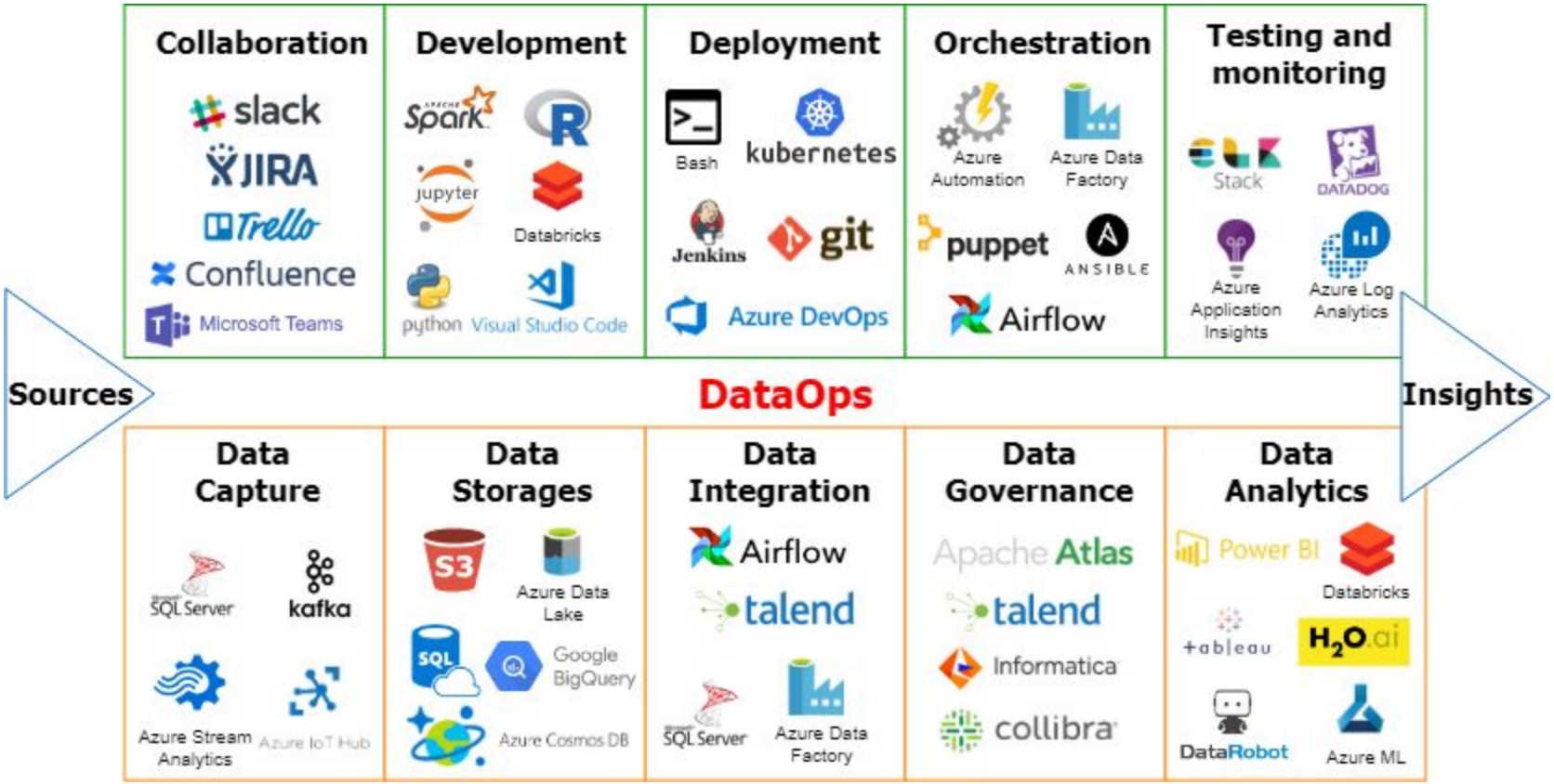
Get familiar with DataOps



Get familiar with DataOps

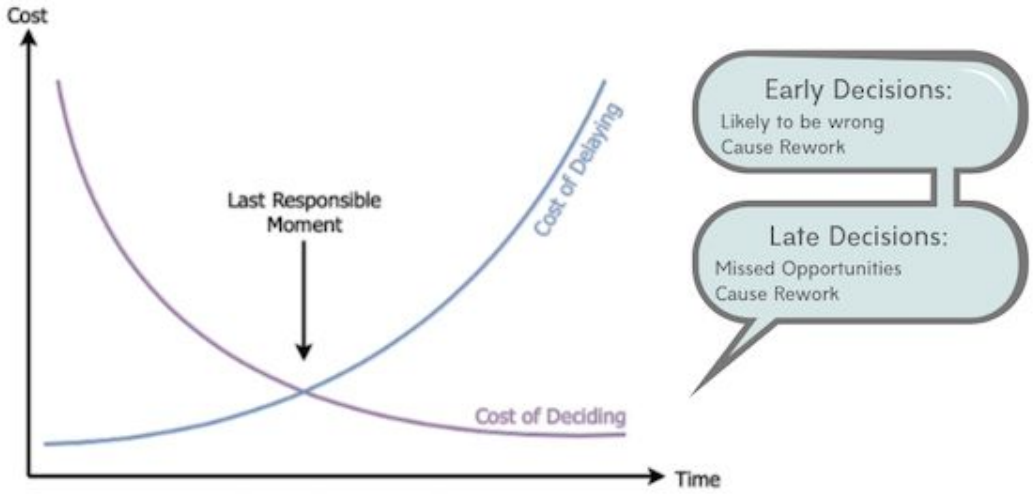


Get familiar with DataOps - Examples



Delay commitments and keep important decisions open

- The principle of **Last Responsible Moment** originates from Lean Software Development
- It emphasises holding on taking important actions and crucial decisions for as long as possible.



Why Last Responsible Moment is important in cloud analytics?

Expect new improvements and upgrades all the time

- 10 January, 2020

  Amazon Redshift at re:Invent 2019
  09:19 PM • AWS Amazon Redshift Analytics

  Lustre File System Performance on Oracle Cloud Infrastructure
  07:13 PM • Oracle Partners Performance

  Oracle Cloud Days 2020: Building Your Future
  06:27 PM • Oracle Events Oracle Cloud Infrastructure

  Exploring container security: Navigate the security seas with ease in GKE v1.15
  05:00 PM • GCP Identity & Security Exploring Container Security

  A home away from home: Wayfair goes hybrid on Google Cloud with 100 Gbps Dedicated Interconnect
  04:00 PM • GCP Networking Customers

- 09 January, 2020

  How Verizon Media Group migrated from on-premises Apache Hadoop and Spark to Amazon EMR
  07:35 PM • AWS Amazon EMR Analytics

  Trying cloud on for size: URBN's Nuuly builds from scratch with Google Cloud
  05:00 PM • GCP Retail Customers

  Exploratory data analysis, feature selection for better ML models
  05:00 PM • GCP Data Analytics AI & Machine Learning

  Join us at Google Cloud Next '20: April 6-8 in SF
  05:00 PM • GCP G Suite Events

- 08 January, 2020

  Put your archive data on ice with new storage offering
  05:00 PM • GCP Storage & Data Transfer

  Creating CI/CD pipelines for ASP.NET 4.x with AWS CodePipeline and AWS Elastic Beanstalk
  04:22 PM • AWS AWS CodeBuild AWS CodeCommit

valdas@maksimavicius.eu
https://www.linkedin.com/in/valdasm/
Twitter: @VMaksimavicius