Red Hat Summit

Connect

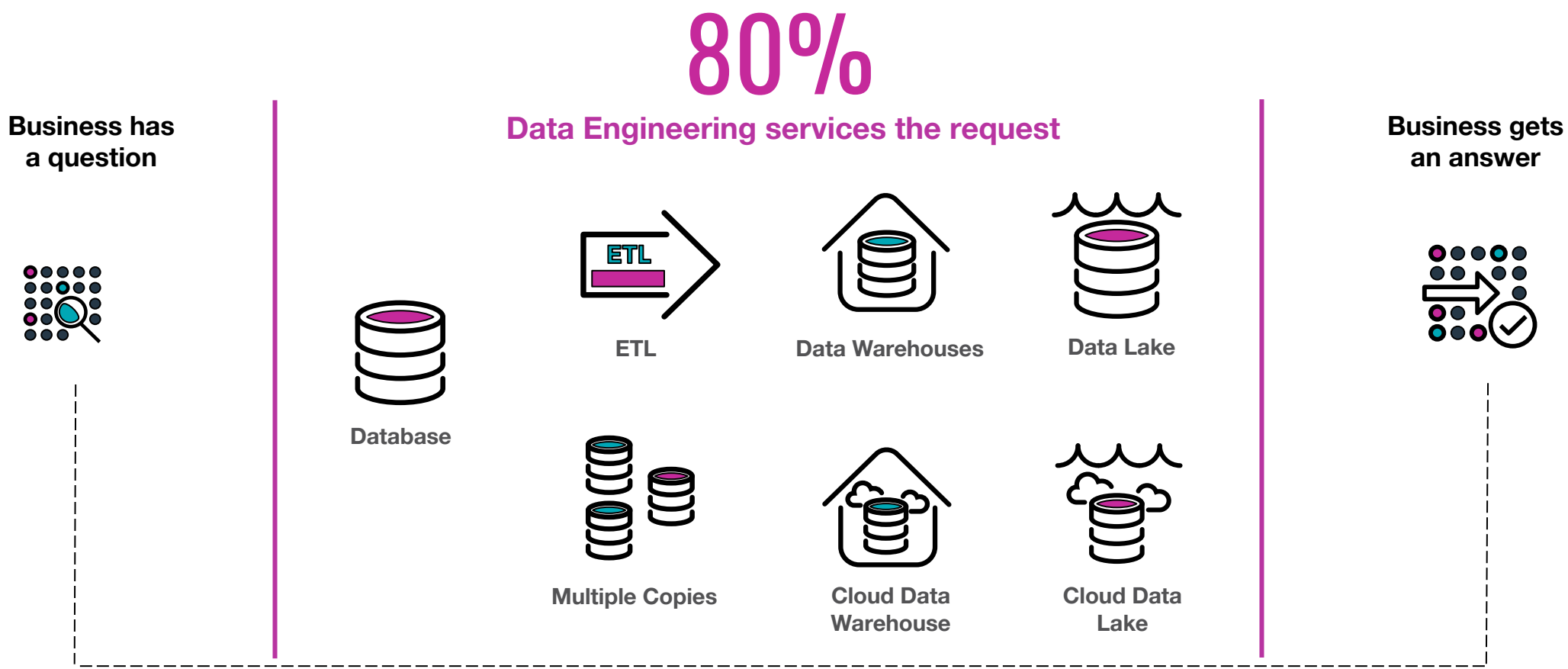
Zukunftssicher und Flexibel

Hybrid Cloud AI und Datenstrategie in der Praxis

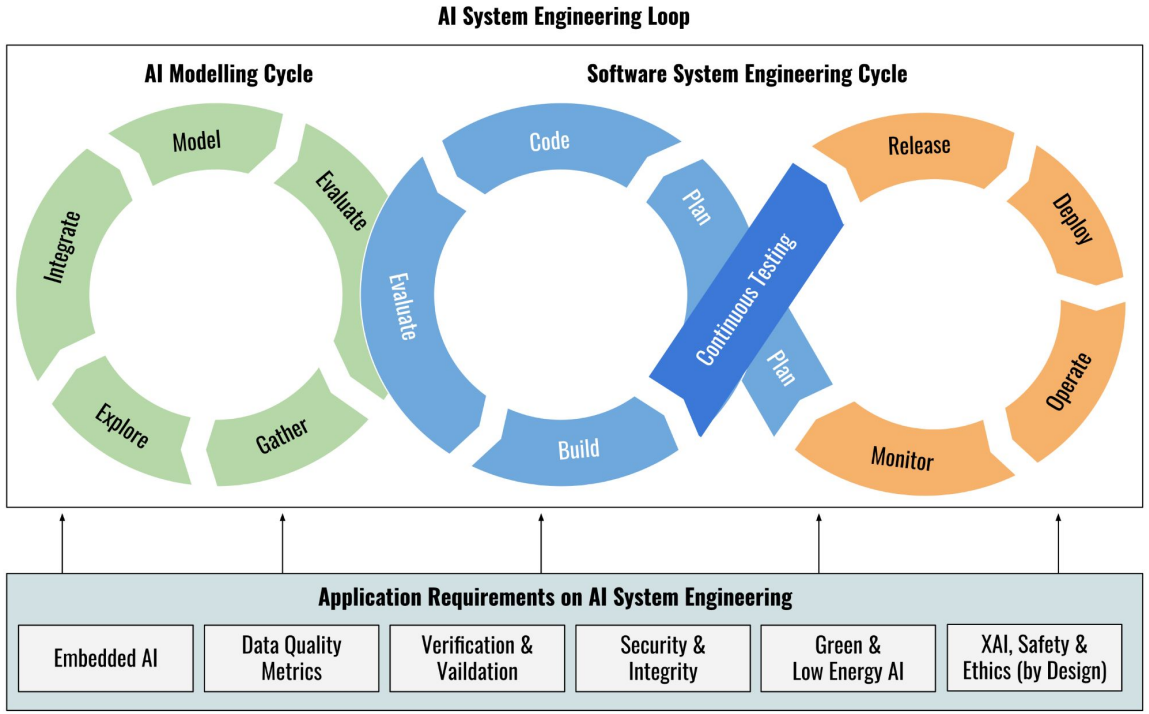
Miriam Bressan & Jonas Janz

Red Hat

Today’s data management approach delays analytics



Delays decision-making, increases data costs & complexity



> Applications which don't use data to build continuous learning systems to scale and adapt autonomously are going to be obsolete.

Mark Little
Chief Data Strategist & Head of Engineering at Anonos

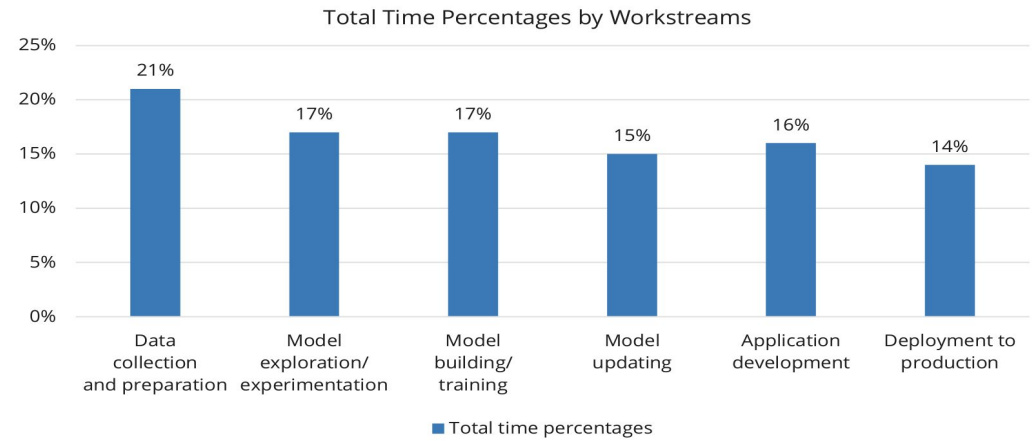
Fischer, Lukas, Lisa Ehrlinger, Verena Geist, Rudolf Ramler, Florian Sobiezky, Werner Zellinger, David Brunner, Mohit Kumar, and Bernhard Moser. 2021. "AI System Engineering—Key Challenges and Lessons Learned" Machine Learning and Knowledge Extraction 3, no. 1: 56-83

The Rise of AI requires better Data Integration capabilities

Data is at the core of AI and also is a key challenge for our customers to successfully deliver on the vision we have for OpenShift AI

Model Development Challenges

Time Spent on Each Stage of the AI/ML Application Life Cycle



Getting data into the platform is a key challenge

According to IDC's AI Strategies View 2021 Survey, the biggest model development challenge that 46% of organizations face is the difficulty to get data into the platform.

More time is spent on data than on data science

Because of diversity of data and data silos, organizations spend the largest percentage (21%) of their total time in AI/ML life cycle in data collection/preparation

AI/ML is driving a huge growth in data platform market

The global automated data platform market size was estimated at USD 1.3 billion in 2022 and it is expected to hit around USD 7.5 billion by 2032 (CAGR of 19.15%)

Sources:
Scaling AI/ML Initiatives: The Critical Role of Data (IDC, February 2022)
Automated Data Platform Market (Precedence Research, 2022)

Red Hat

AI is becoming ubiquitous

AI technology is turning into a competitive advantage

Government

Smart City
Sensor-based asset monitoring

Manufacturing

Quality assurance

Retail

Digital in-store experience

Health–life science

Patient diagnosis/treatment

Energy

Monitoring and control

Automotive

Autonomous driving
Predictive maintenance

Financial Services

Fraud detection
Risk analysis

Telecommunications

Threat detection

Insurance

Automated claims processing

Red Hat

It’s all about the models – or not?

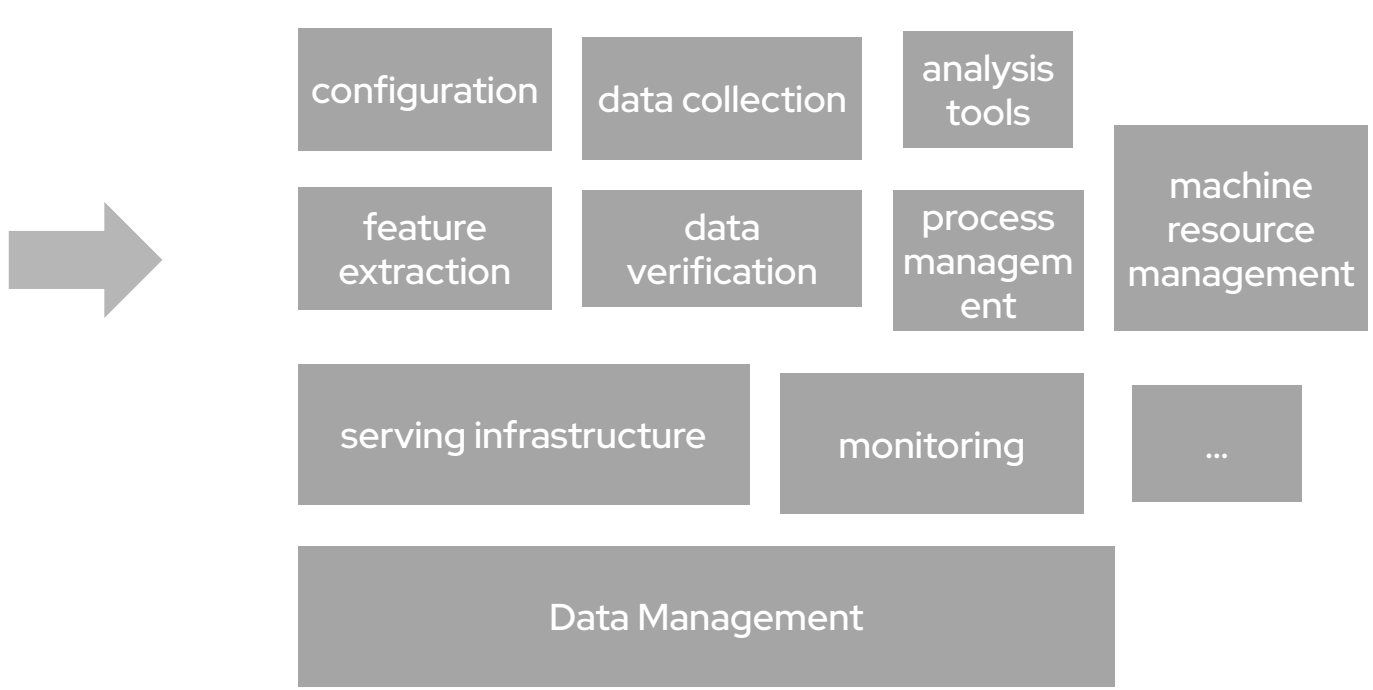
Lack of focus on end-to-end system builds technical debt

Experimentation

Model Code

Production

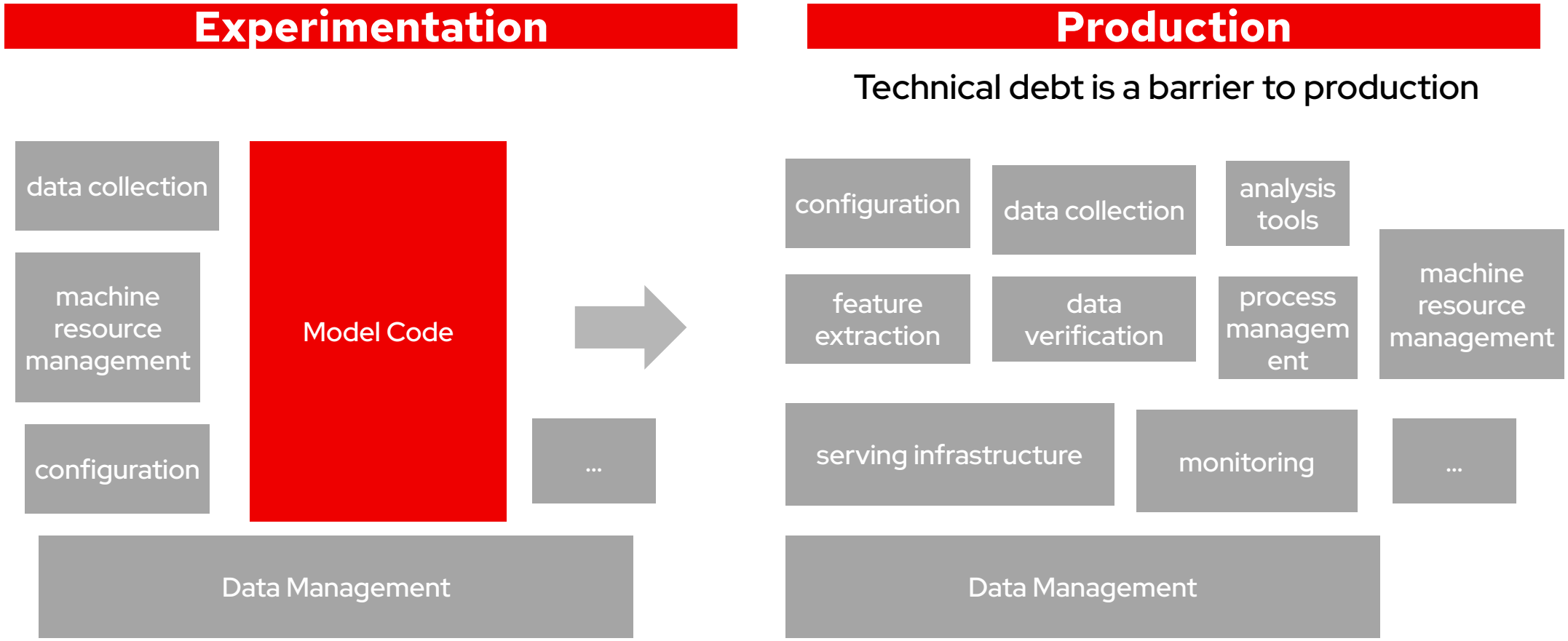
Technical debt is a barrier to production



Red Hat

It's all about the models - or not?

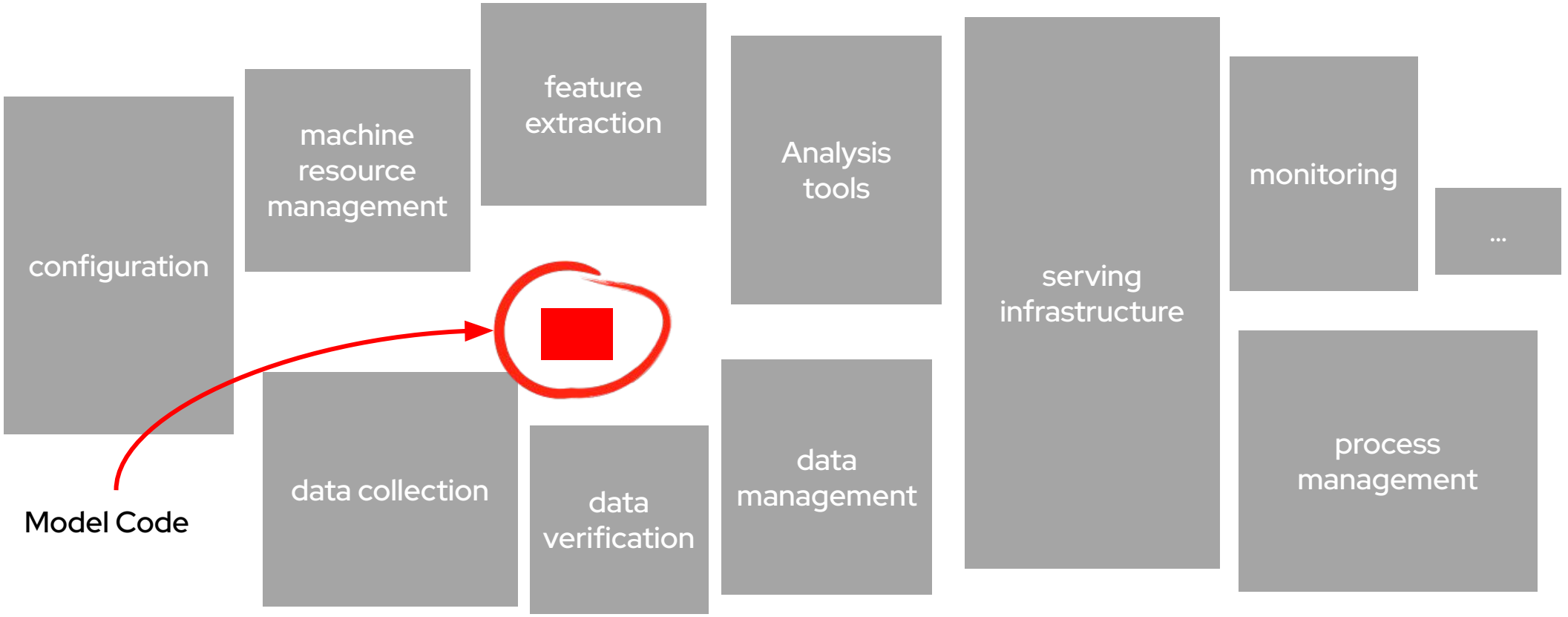
Lack of focus on end-to-end system builds technical debt



Red Hat

Deploying intelligent applications using ML is not trivial

The model is a small portion

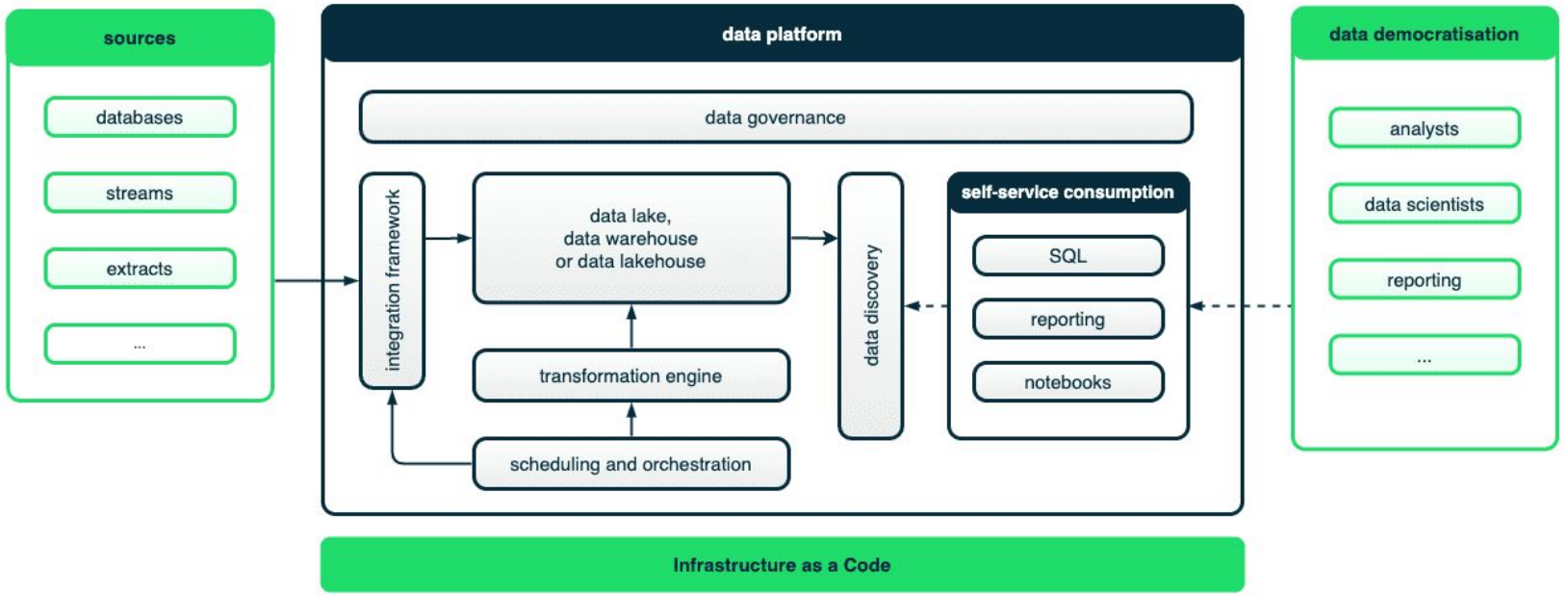


(Adapted from Sculley et al., "Hidden Technical Debt in Machine Learning Systems." NIPS 2015)

Source: https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

Here enters the Modern Data Stack

Cloud offerings in the AI/ML space are backed by data platforms to ease adoption



Emergence of AI Foundation Models requires data management guardrails

Enterprises have increased need for reverse ETL patterns

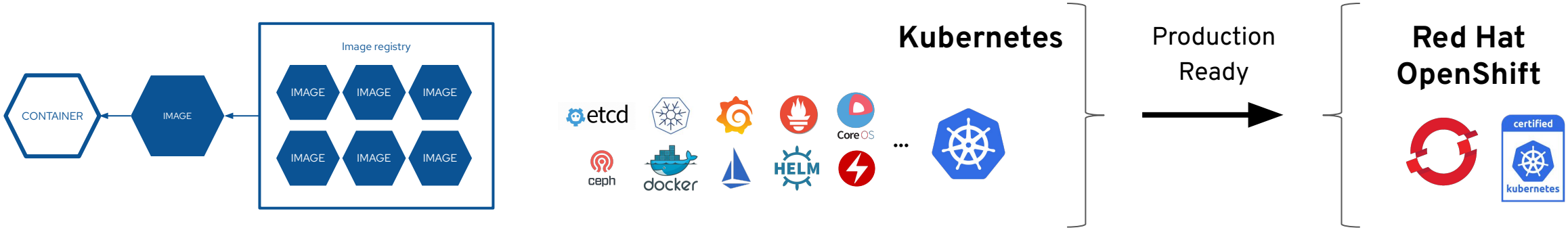
Hyperscalers are investing in Zero ETL capabilities to enable connected data workflows

Source:
6 Key Trends in Shaping the Modern Data Stack in 2023 (Akash Rathi, Anurag Gupta for Elevation Capital, 22 March 2023)
Modern Data Platform - the what's, why's and how's? (getindata on 29 June 2022)
Data Management Implications for Generative AI (Datanami, 25 May 2023)

Hybrid Cloud is NOT just about Infrastructure

Applications	Infrastructure	Development, IT processes, and skills
AI/ML Data/Analytics Serverless Cloud-native and microservices Java™ .Net ISV	63% of organizations are already using hybrid cloud today.[1] 54% of those not using hybrid cloud today plan to within 24 months.[1] Bare metal Virtualization Private cloud Edge Public cloud	Developer tools Pipeline and processes People and policies

Why containers, K8s, OpenShift for Data Science?



Why containers?

- Fewer resources
- Environment isolation
- Quick deployment
- Quick startup/shutdown
- Encapsulation and portability
- Reusability
- Reproducibility

Why Kubernetes?

- Automated rollouts and rollbacks
- Self-healing
- Service discovery and load balancing
- Horizontal scaling
- Designed for extensibility

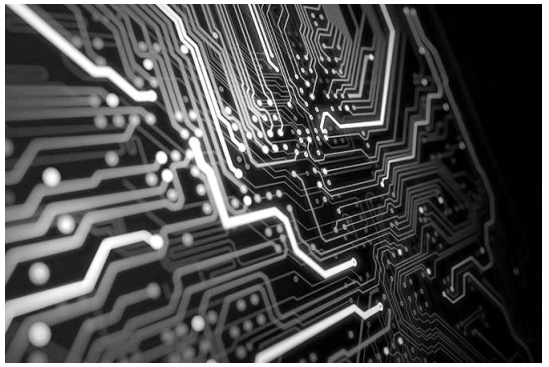
Why OpenShift?

- Cloud and Infra Agnostic
- GPU Support
- Multi-tenancy
- Zero Trust Security Model
- Metrics and Monitoring
- IAM integration
- Web UI based Workflows

Red Hat

Capabilities of a Modern Data Stack

Integrates with your Intelligent Application Platform for DataOps and Data Governance



Data Federation

Work with ecosystem to provide readily available hybrid deployment architecture for integration of distributed heterogeneous data sources, with dynamic cross-cluster query routing

DataOps

Build up an offering based on open source tooling for managing the whole data lifecycle, from creation to collection to processing to monitoring that data, through a data as code approach

Data Governance

Provide a federated data governance capability encompassing Metadata Management, Data Catalog, and Data Policy Management (solving potentially data sovereignty issues)

Further information in the Technical Alignment Document for the Modern Data Stack

But what if your team could also....

Data Engineers

Reduce complexity of existing system

- Faster, more reliable, and scalable queries
- Reduce dependence on ETL pipelines
- Ensure data security + governance

Data Analysts

Quickly develop meaningful insights from your data

- No waiting for data access
- Reduced reliance on data engineering
- Seamlessly join data
- Self-service with BI tool of choice

Line of Business

Increase profitability, growth and revenue

- Speed time-to-market
- Improve customer experience
- De-risk business decisions

Part 1: Self-Service Data Infrastructure

Data Scientist

Data Engineer

Common challenges

- Data availability is restricted
- Long time to provision infrastructure
- Hard to collaborate across technical teams

Benefit

Faster Time to Value

1 — 2 — 3 — 4 — 5

Identity Management

Manage user profiles through federated yet consistent identity management

Role-Based Access

Unified role-based access control to infrastructure & environments

Infra Provisioning

Self-service provisioning and sharing of infrastructure resources

Shared Repository

Enable distributed team collaboration on version-controlled data science projects

Run Data Pipelines

Run a data processing pipeline assembled from python notebooks and scripts

Part 2: Build a data ingestion pipeline as code

Data Engineer

Data Quality Engineer

Common challenges

- Manual operations are prone to user error
- Business logic and data are hard to reproduce
- Lack of transparency and trust in data

Benefit

More Trustworthy Data

1 — 2 — 3 — 4 — 5

Data Extraction

Access controlled data source for automated data extraction

Version Control

Create a repository and implement data version control with Pachyderm

Data Loading

Load data from versioned source into Trino, a federated SQL Engine

Data Transformation

Process and transform the data with data-as-code source control using DBT

Metadata Ingestion

Automatically ingest metadata and build a catalogue with OpenMetadata

Part 3: Manage domain data as a product

Data Scientist

Domain Owner

Common challenges

Inefficient copying of data to enable analysis

Inconsistent security for data in transit

Hard to collaborate across data and knowledge domains

Benefit

Secure and Unified Data Resources

1

Discover Data

Discover data sources in a centralized data catalogue

2

Access Control Management

Centralized security framework to manage fine grained access control

3

Heterogeneous Queries

Single point of access and consumption for federating heterogeneous data

4

Federated Governance

Enabling secure queries, consistent governance and auditing for compliance

5

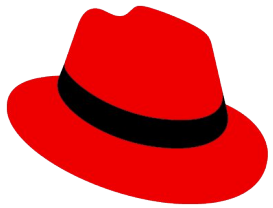
Use Data

Build custom visualization and interactive dashboards to share data

Scan QR Code - Trust me, it's safe 🤓

Bist du interessiert an einem Deeper Dive Webinar?

https://forms.gle/t3xF2iZj2dTWN8rP6



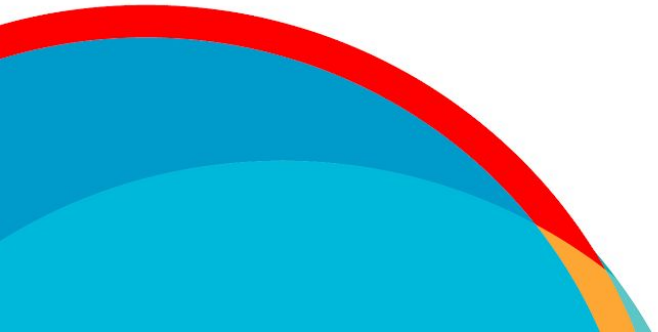
Red Hat

Miriam Bressan

miri@redhat.com
linkedin.com/in/themiri

Jonas Janz

jjanz@redhat.com
linkedin.com/in/jonas-janz/



Session: 17:25 - 17:55

Jetzt Session bewerten!

Einfach QR-Code scannen, Session wählen und bewerten.
Vielen Dank!

red.ht/rhsc24-de-s8

Red Hat

Red Hat Summit

Connect

Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat