July 11, 2023

IDMC: The Essential Toolset for Data Scientists - Encompassing Machine Learning, Machine Learning Ops, and Data Engineering

- Nirav Joshi, Senior Technical Manager, IPS

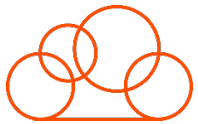
Informatica

Housekeeping Tips

- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our INFASupport YouTube channel and Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

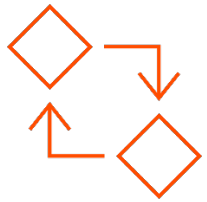
Informatica

Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

Informatica

More Information

Success Portal

https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com

University

https://www.informatica.com/in/services-and-training/informatica-university.html

Informatica

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Informatica

Agenda:-

- INFA CORE landscape in IDMC Platform
- INFA CORE Overview and API
- ModelServe Architecture in IDMC Platform
- ModelServe Overview and usage by personas

Challenges

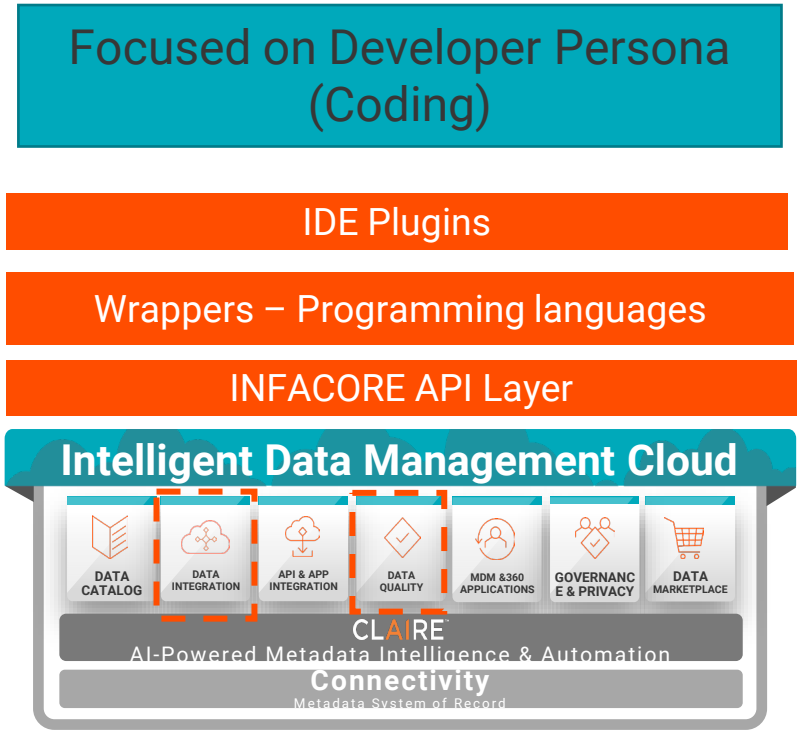
No code users	PERSONAS	Low code users

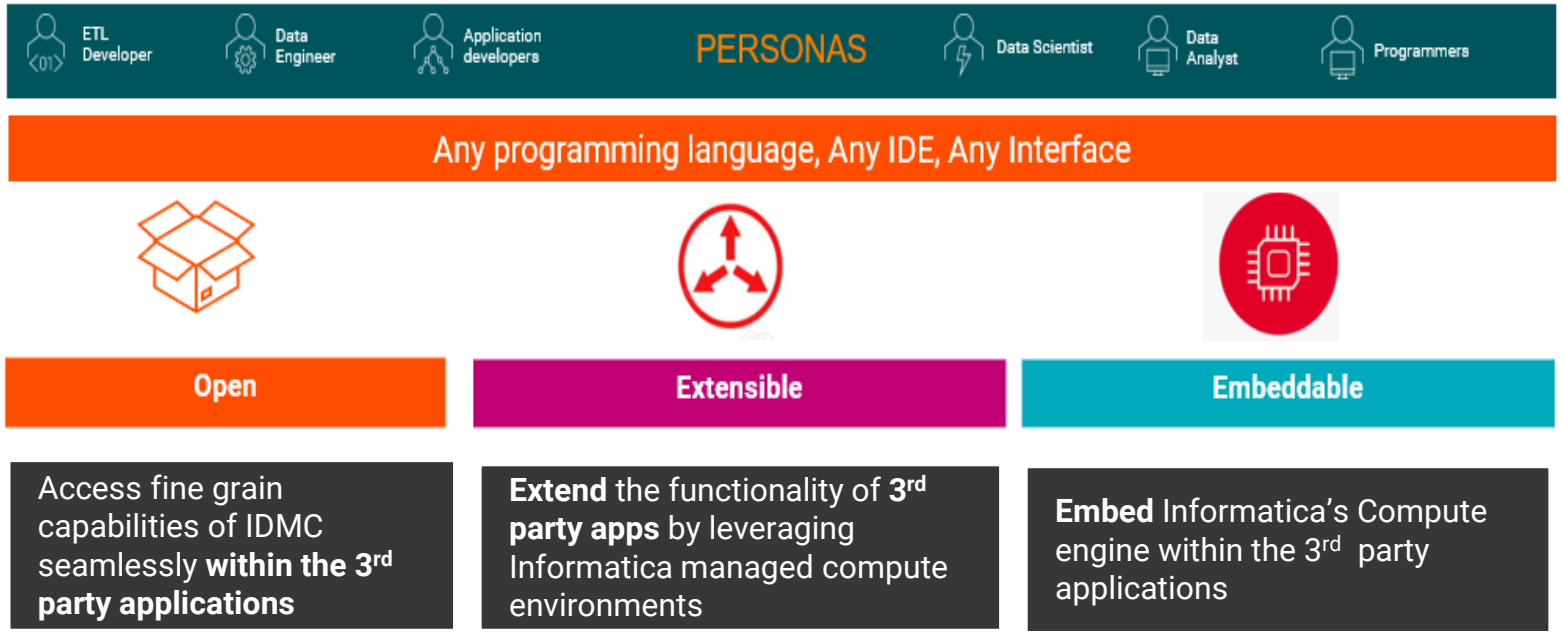
Key Challenges	**Productivity** **Maximum reusability** **Rapid prototyping** **Efficiency** **Built in governance & documentations** **Ease of maintenance** **OOTB Options** **Lot of Context switching/integrations**

No pervasive platform to support the needs of all data personas in a data driven enterprise

INFACore - Overview

An open , extensible ,embeddable framework for democratization of data management for all developers and all data driven applications.





Use cases	
Direct	In Direct
Used directly within the program	Embedded within the application/product (OEM)

Informatica

What Is Apache Arrow?

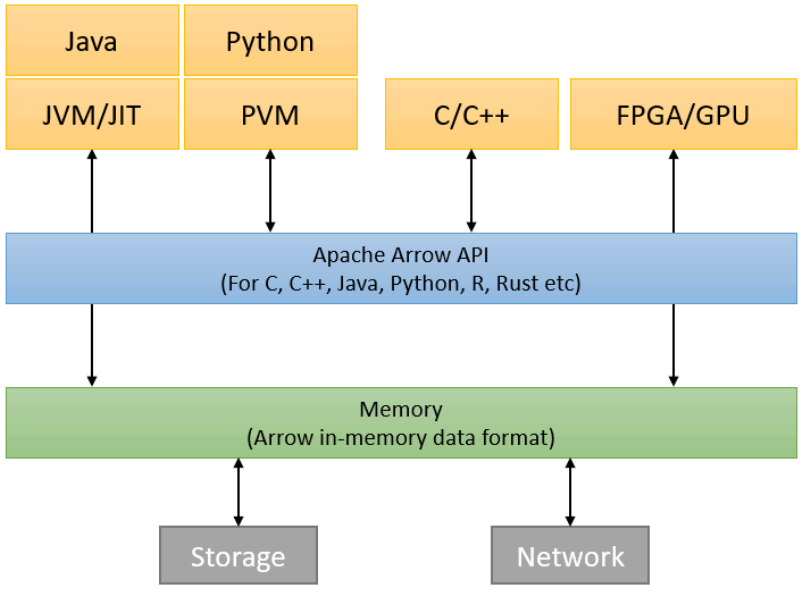
Apache Arrow improves the speed of data analytics by creating a standard columnar memory format that any computer language can understand. In addition to being a development platform, it also provides libraries for software.

Apache Arrow performance allows for the transfer of data without the cost of serialization (the process of translating data into a format that can be stored). Apache Arrow is a standard that can be implemented by any computer program that processes memory data.

How Does Apache Arrow Work?

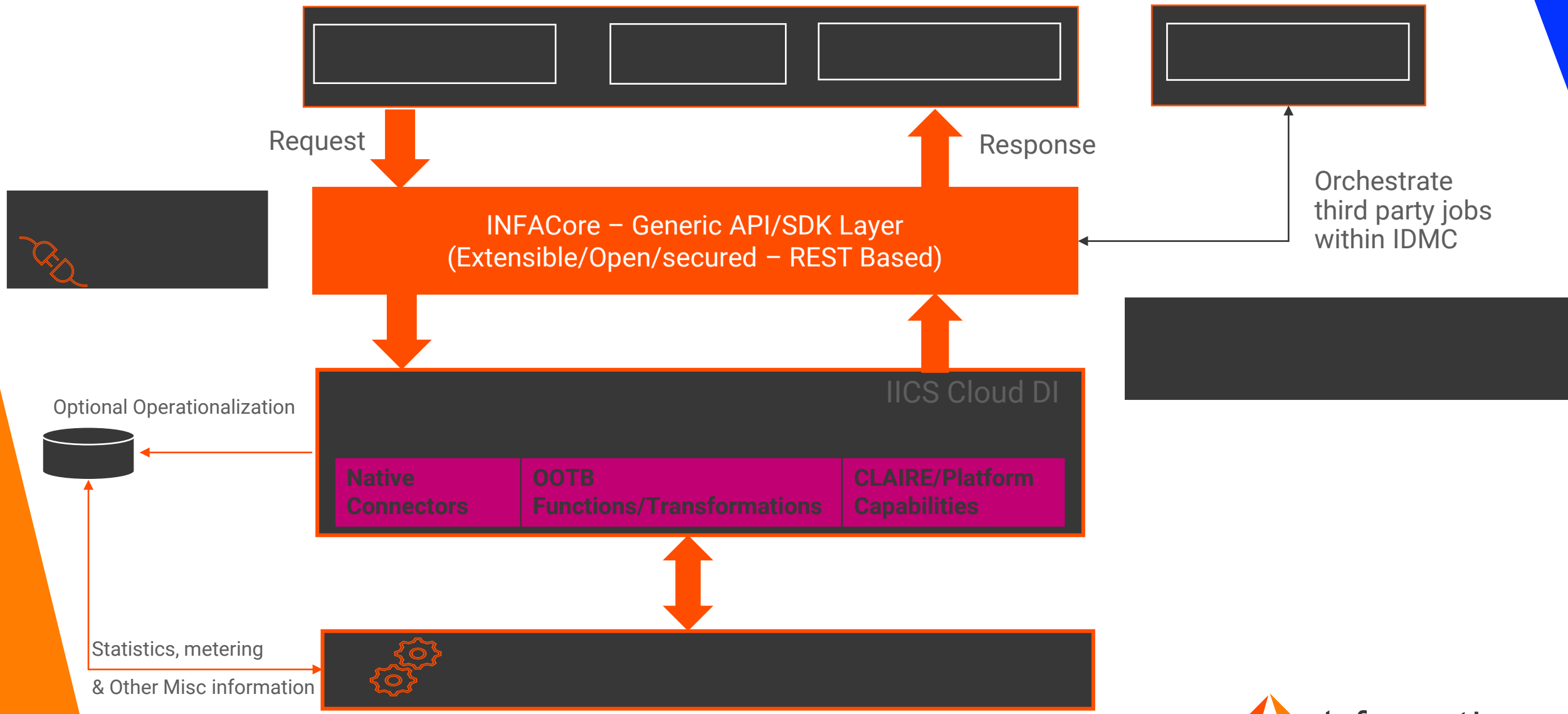
Apache Arrow acts as an interface between different computer programming languages and systems. By creating a standard for columnar data layout (versus rows) for memory processing, it speeds up the transfer of data by eliminating unnecessary input/output communication. It also provides caching benefits for data structures. This optimizes the performance of modern central processing units (CPUs) and graphics processing units (GPUs).

Apache Arrow processes large amounts of data quickly by using Single Instruction Multiple Data (SIMD). Sets of data are broken into batches that fit the cache layers of a CPU. The Apache Arrow project has a standard format allowing for seamless sharing of data between systems instead of using CPU cycles to convert data between formats.



Informatica

High Level Architecture – INFACore



Informatica

DSML and MLOps

DSML and MLOps

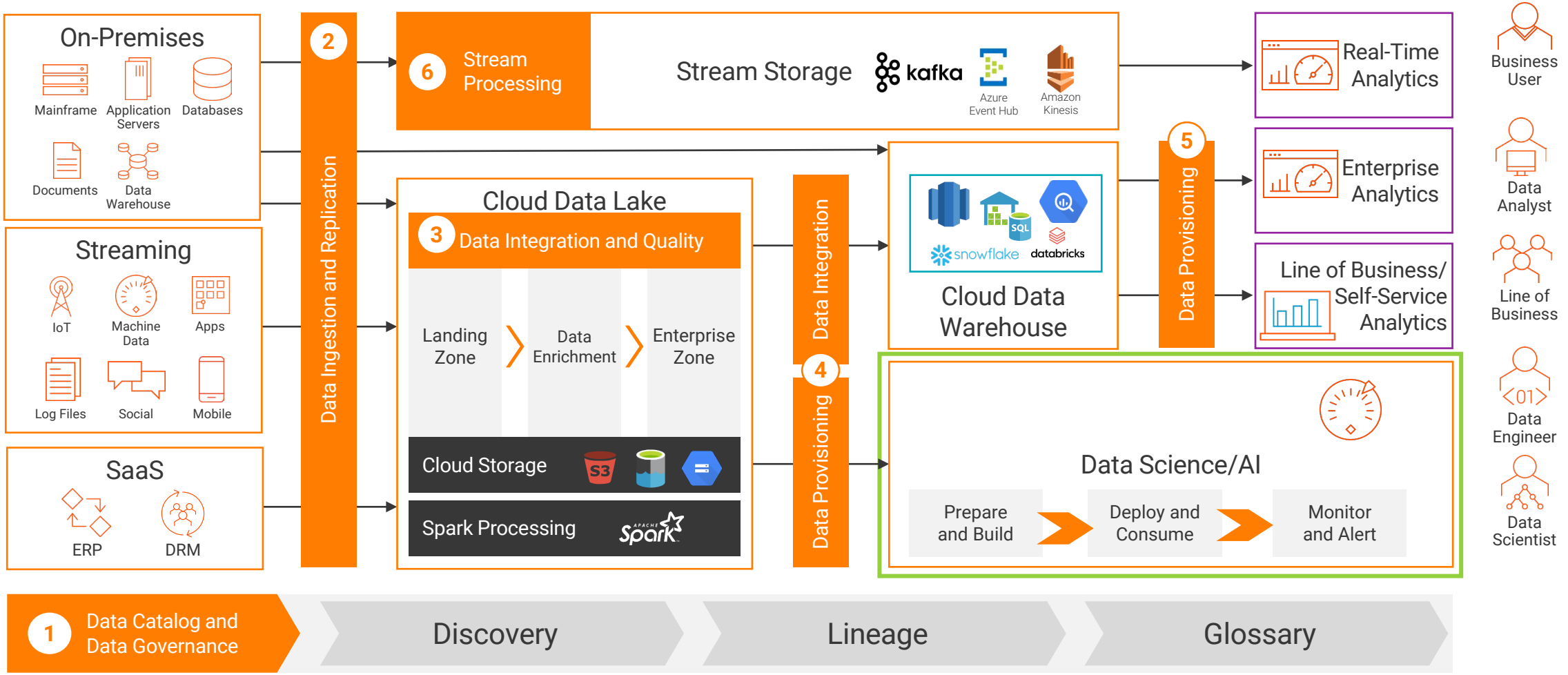
DSML—Data Science and Machine Learning

- A DSML platform provides support for a portfolio of products, components, libraries and frameworks that are coherently integrated. These could be proprietary, partner or open source. The primary audience for this platform includes data science professionals such as expert data scientists, citizen data scientists, data engineers and ML engineers/specialists.

MLOps—Machine Learning Operationalization

- MLOps allows for the implementation of a complete pipeline that supports the continuous integration and continuous delivery of machine learning models in a production environment. This enables the operationalization of the end-to-end process.

Where Do We See DSML in IDMC?



Machine Learning Use Case

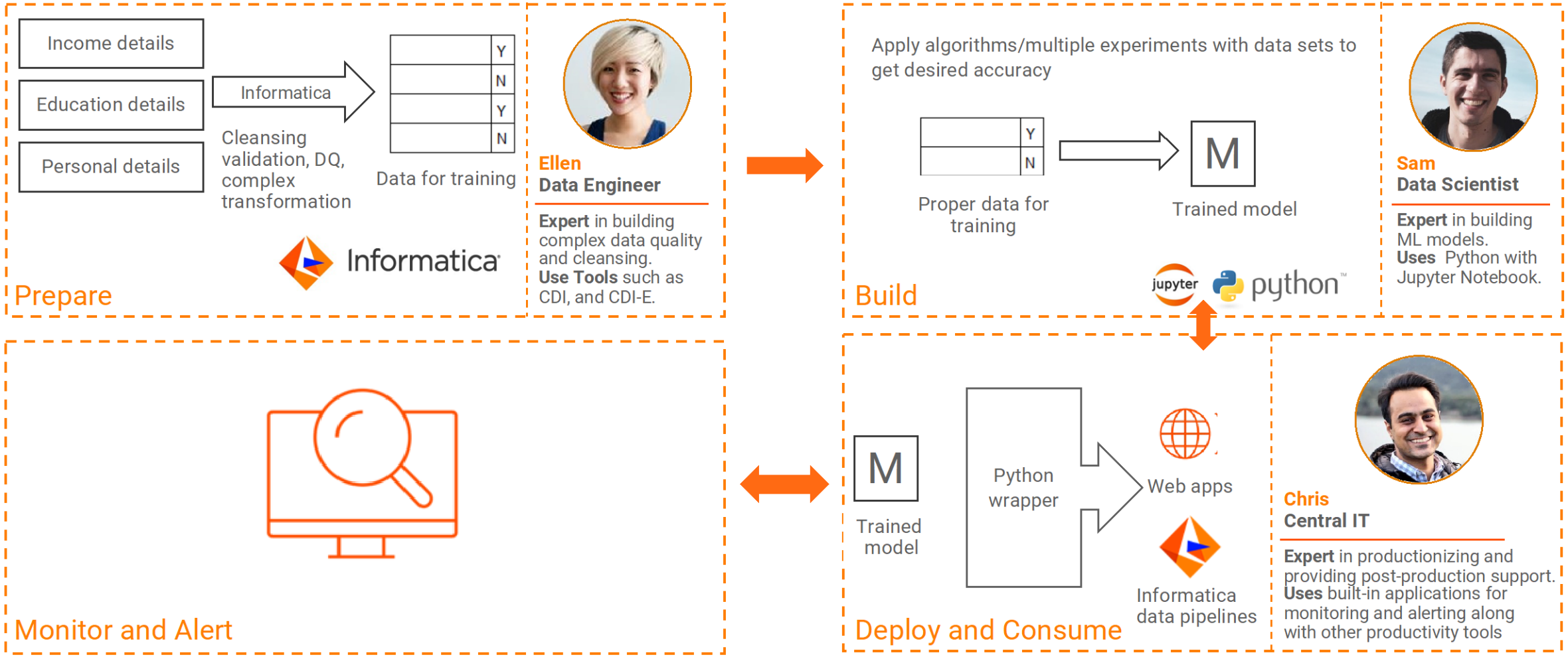
Scenario

- ABC Financial Services Corp. is a specialist in loan/mortgage services. They need to validate loan applications for potential default. This is an extremely slow and complicated process.

Goal

- Use machine learning (ML) to augment the static rules to find out potential loan defaults
- Use the historical loan information to develop/train the ML model which can be used to predict potential default based on the loan application details
- The prediction of the ML model will be available as one of the key metrics that will be used to decide whether to give a loan or not
- This ML model will be used by multiple applications such as:
  - Informatica mapping for batch processing
  - Java-based web application

Naive Solution



Other Challenges

User Personas

- Skillset gap
- Analyst transitioning to citizen data scientists
- Lack of coordination between different project stakeholders

General

- Debugging is hard
- Diverse frameworks and algorithms
- Uncertain ROI
- Difficult to monitor and alert proactively

Governance

- Compliance and regulatory needs
- Audit
- Fairness and explainability
- Access control
- Versioning
- Experiment tracking

AI/ML Taking Mainstream

Require automation

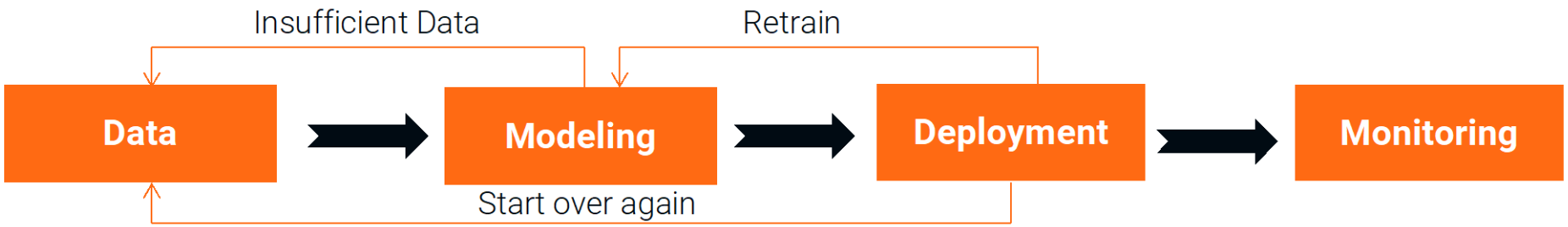
- CI/CD pipelines for models

Data

- Varied sources
- Exabyte scale
- Quality

Lack of best practices and standards

How Do We See MLOps at Informatica?



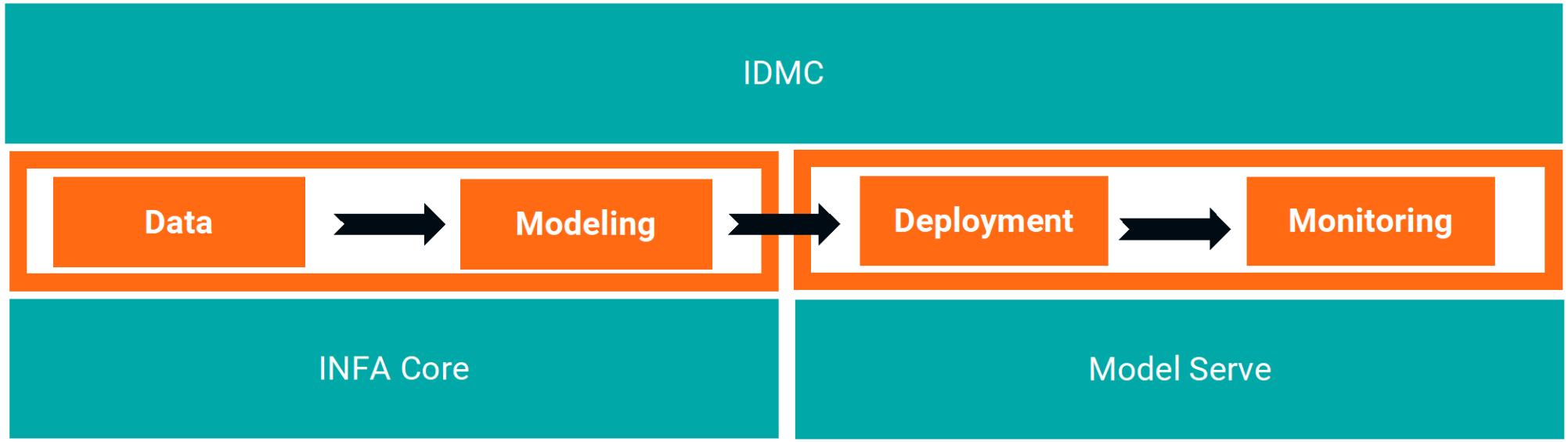
- Define data
- Procure data
- Data cleansing
- Data preparation

- Model selection
- Model training
- Error analysis

- Model deployment
- Model consumption

- Model performance
- Alerts
- KIP validation

Informatica's Solution to MLOps



Model Serve

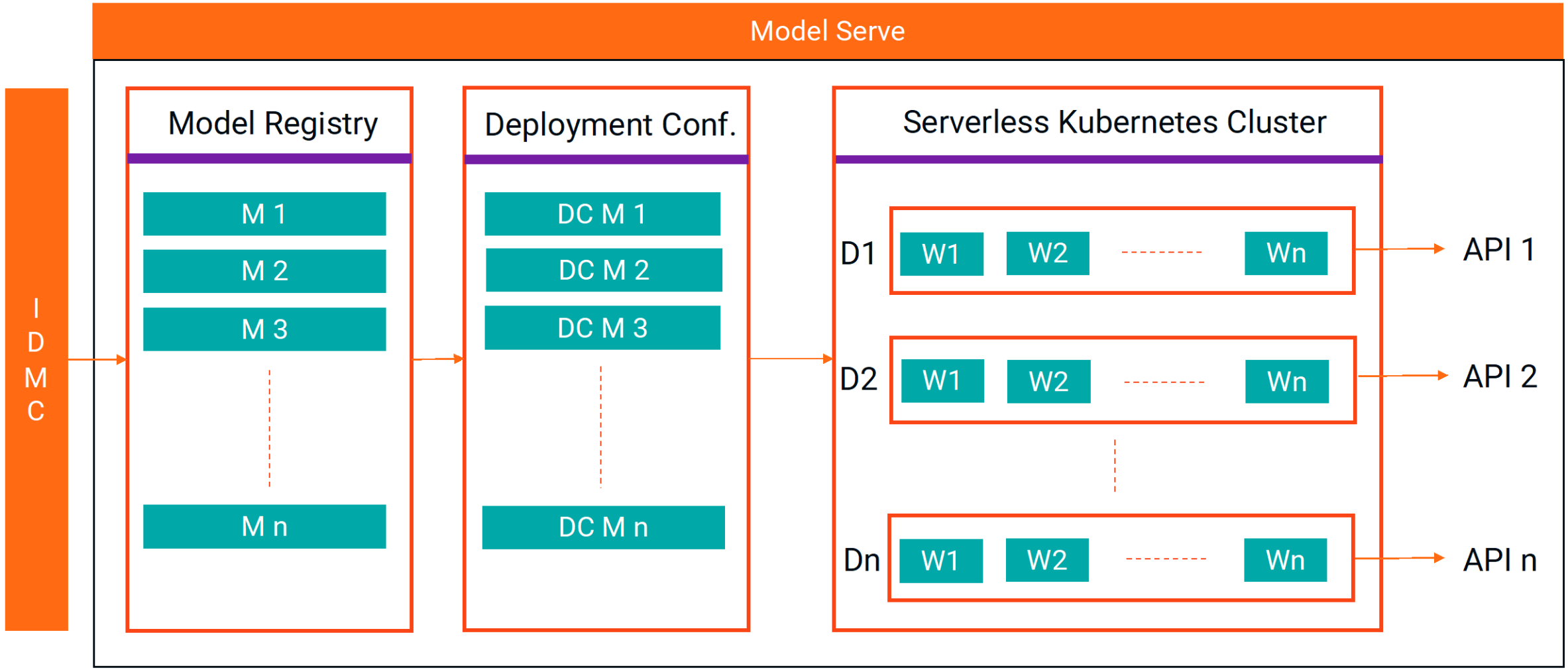
Key Highlights

- Centralized model registry and management within IDMC
- Framework/language/compute agnostic platform for operationalizing any model at scale
- One-click elastic serving of registered ML models
- Monitoring and alerting of served models
- Access controls and integrated governance

Benefits

- Operationalization of ML models at scale
- Ease of maintenance of ML lifecycle
- Integrated dataOps for time delivery of trusted data

Model Serve Architecture

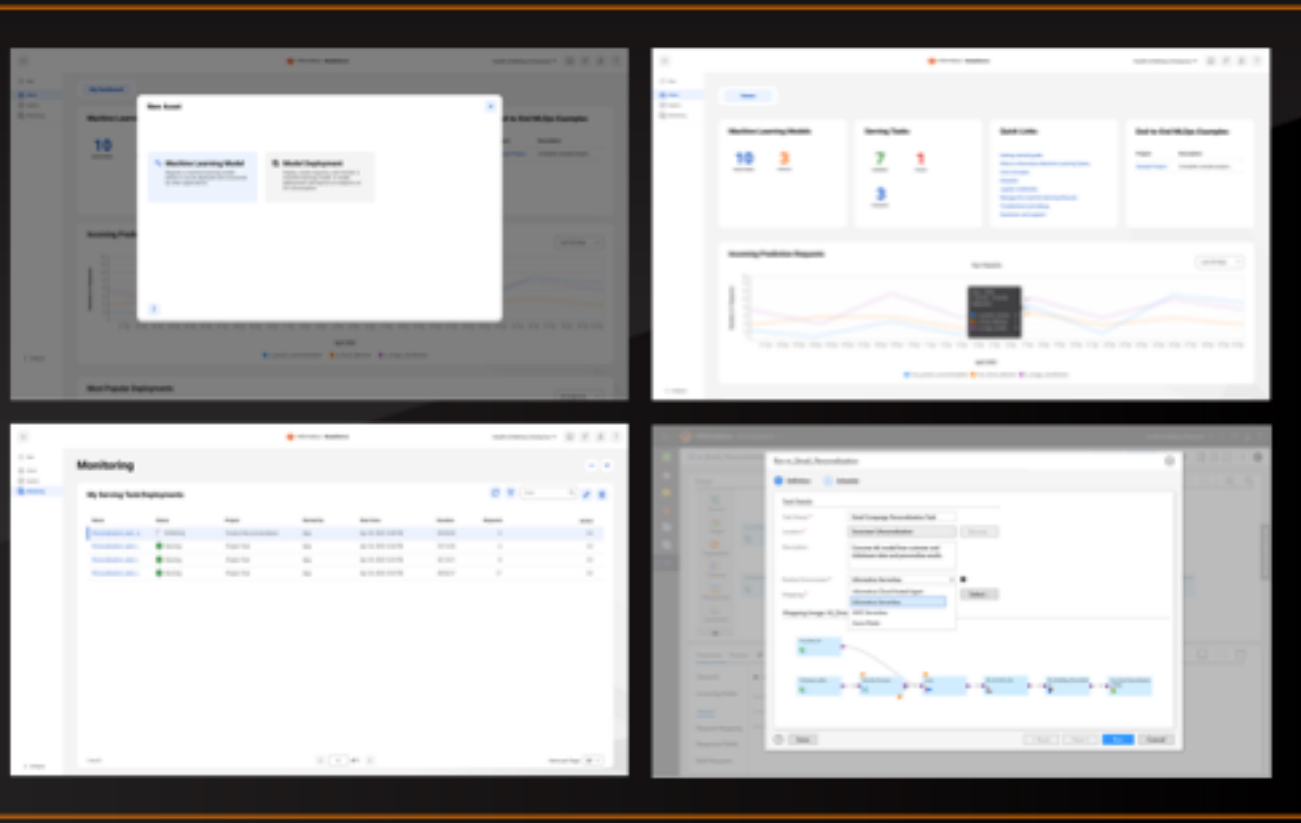


RESOURCE CONSTRAINT | COST | COMPLEXITY

One-click, serverless deployment of ANY AIML Model

Only with Informatica, data scientists and ML engineers can operationalize AI/ML models @ scale with ModelServe

- Simple, easy-to-use, wizard-driven approach for data scientists and ML engineers to deploy and operationalize any AI/ML models at scale
- Provide flexibility for data scientists and ML engineers to build their AI/ML models in any framework and consume them in any application
- Enable data scientists to accelerate AI/ML initiatives with high-quality, trusted, and governed data
- Improve productivity of data science teams by streamlining and automating the process of building, deploying, and monitoring machine learning models
- Enhance model performance with timely delivery of trusted data using integrated DataOps



Sign up for private preview
https://bit.ly/infa22preview

Informatica