WELLS FARGO

Multi-channel multi-grain forecasting on high velocity data

Open Data Science Conference - Asia Pacific 2024

Thejas Bhat
Naveen Rathani

Meet the speakers

Thejas Bhat

Senior Data Scientist with over five years of industry experience. He is passionate about opportunities in responsible decision making by drawing business insights from data, developing deployable solutions, and contributing to significant positive impact of the organization. Thejas is a strong engineering professional with a bachelor's degree focused on electronics and communication engineering and a Masters of Management in Business Analytics from Indian Institute of Science. Thejas started his career as software engineer in the information technology and service industry.

Naveen Rathani

Data Science Manager with close to 12 years of industry experience. Proficient in deploying complex machine learning and statistical modelling algorithms, Naveen has a knack for solving real world business challenges using data and analytics. Previously, Naveen worked in retail, telecom, and insurance verticals for the European and North American markets. Naveen graduated from Birla Institute of Technology And Science, Pilani with a bachelor’s degree in chemical engineering and a master's in mathematics. He is currently pursuing a management degree from Indian Institute of management Bangalore.

How much data do organizations really deal with today?

- **Dynamic pricing and inventory management**
  - Global e-commerce sales are projected to reach $6.54 trillion in 2024 [1]
  - Amazon processes over 400 orders per second during peak times, with millions of transactions per day [2]
  - **Forecasting needs:** Manage inventory levels, predict demand, and adjust prices dynamically
  - **Parallels:** Online advertising | payment's liquidity and fulfilment
- **Network traffic management and service reliability**
  - Global mobile data traffic averages ~50 EB / month in 2024 [3]
  - AT&T handles >204 petabytes of data traffic on an average business day [4]
  - **Forecasting need:** Manage network congestion and ensure service reliability.
  - **Parallels:** Web-traffic |digital footfalls | online payment operations
- **Delivery of energy and utilities related services**
  - Global smart grid market size is over $60 billion as of 2024, with significant data generated from smart meters & sensors. [5]
  - Smart meters can generate data every 15 minutes, leading to billions of data points annually for a single utility. [6]
  - **Forecasting need:** Balance supply-demand, ensure grid stability, and integrate energy generation and consumption sources in real-time.
  - **Parallels:** Esports | wearables | healthcare | manufacturing

Imaging monitoring suspicious activity on your digital channel which sees 100 million visits each day

Real-time demand forecasting

Forecasts that give a multiple level view

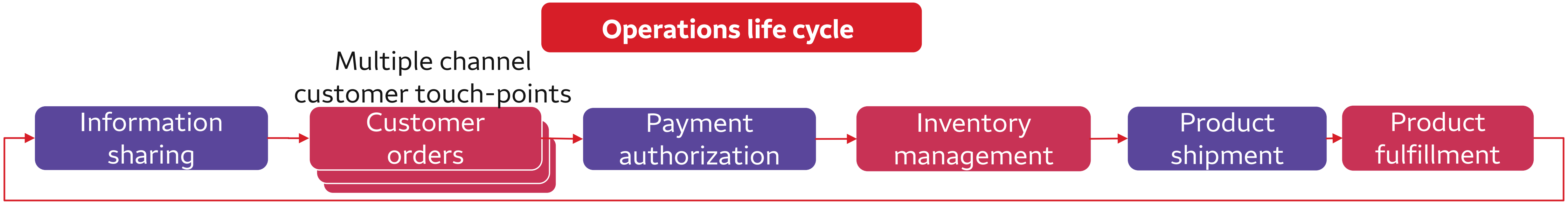
Forecasts at multiple time grains

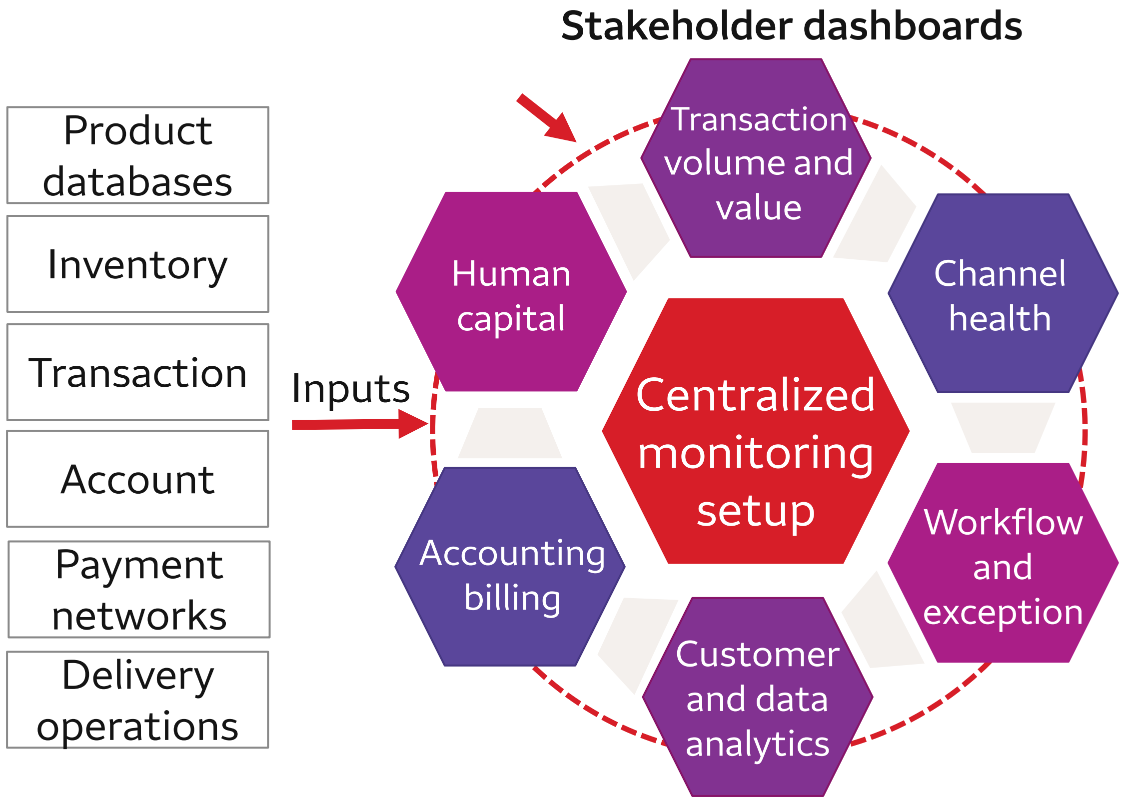
Forecasts that are dynamic

Forecasts that reconcile

Forecasts that update real-time

Operations in a high-velocity digital business involve various stages





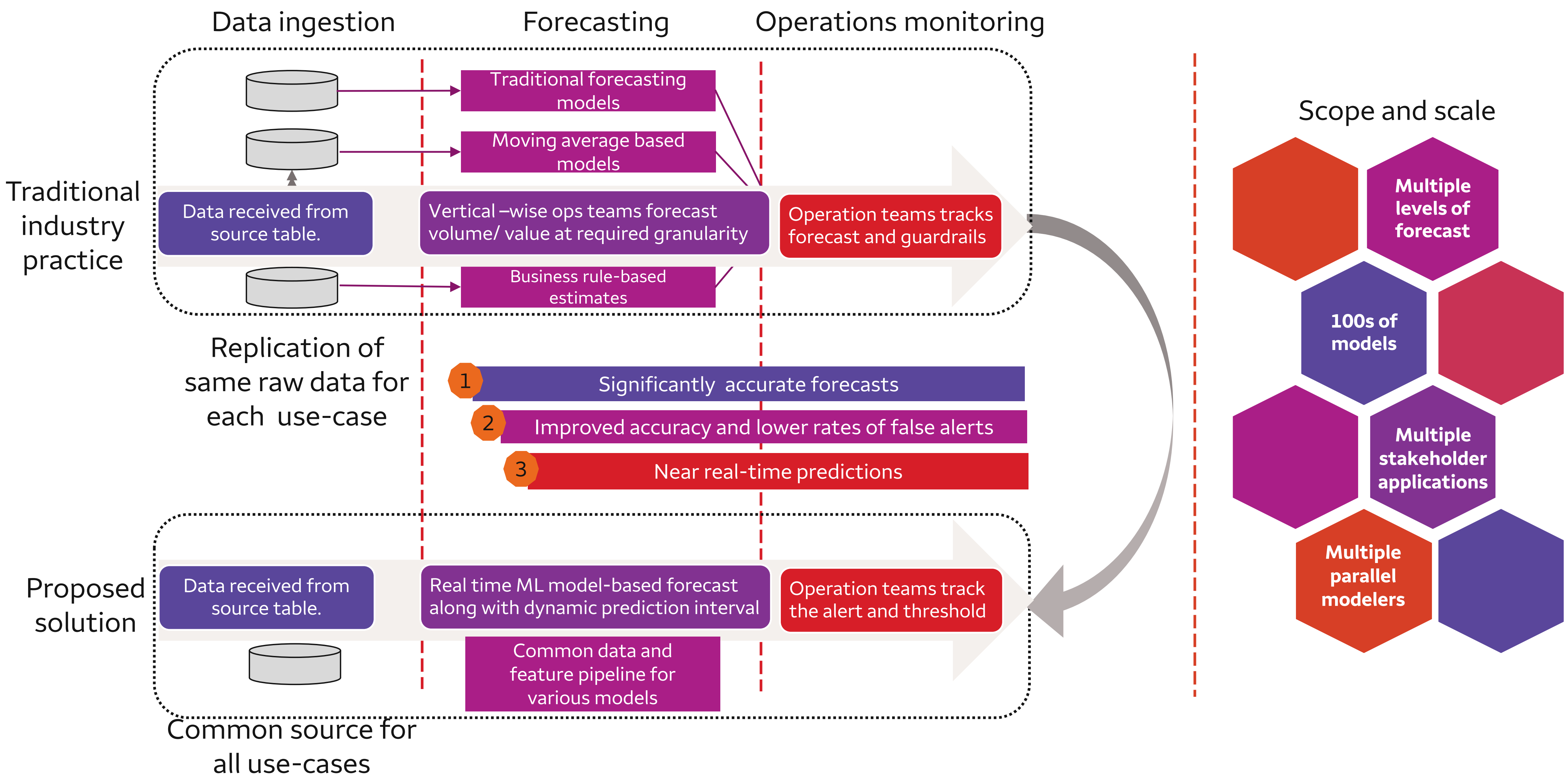
Any company operations team's target state:

- Seeks to aggregate data across a comprehensive set of inputs
- Enables end-to-end operation monitoring, proactive alerting, and pre-emptive control actions
- Improves efficiency, enhance customer experience and
- Maintains competitive advantage

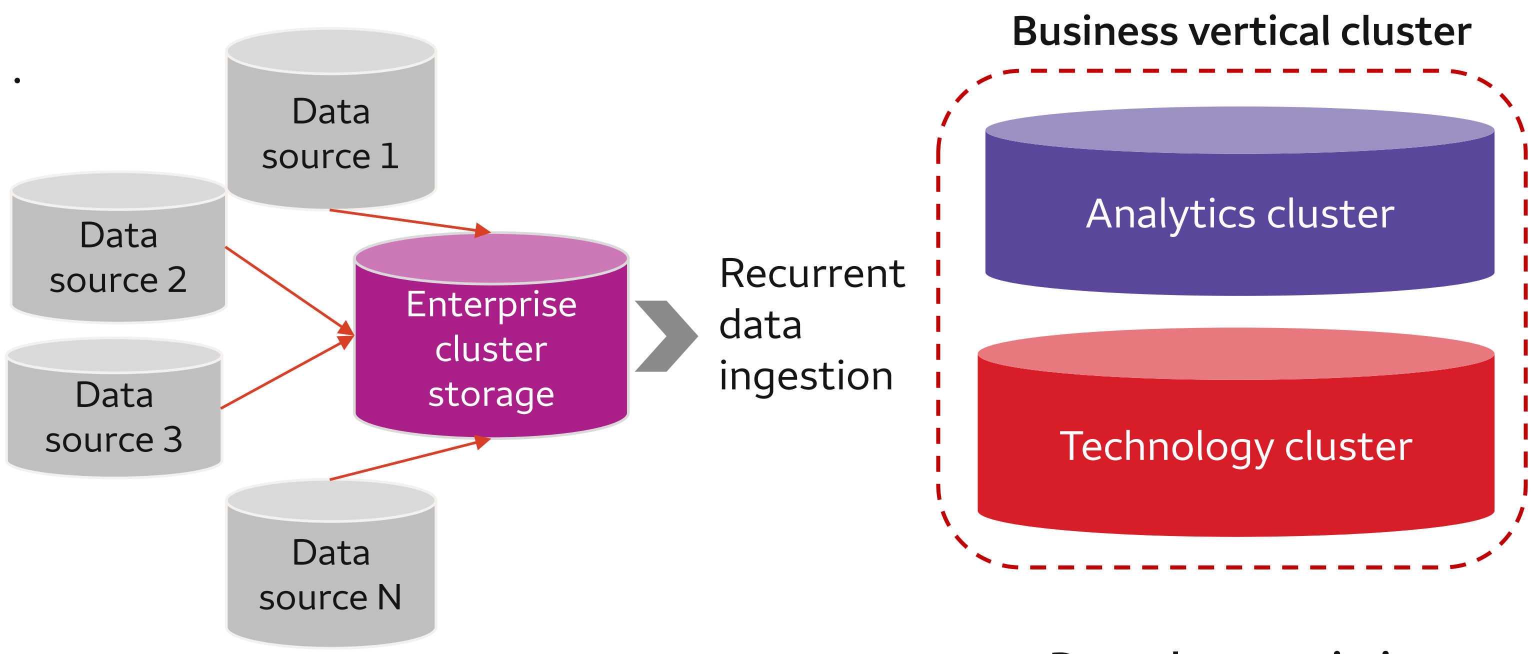
The why, what, and how?

Leveraging advanced analytics and machine learning techniques,	…supporting target state operations program and data strategy,	…using cutting-edge technology and platforms.
• Forecast transaction value and volume • Detect anomalies	• Manage operations • Identify areas of operational efficiency • Pre-empt future trends	• Automated machine learning experimentation framework • Distributed computing platform • Real time data streaming services

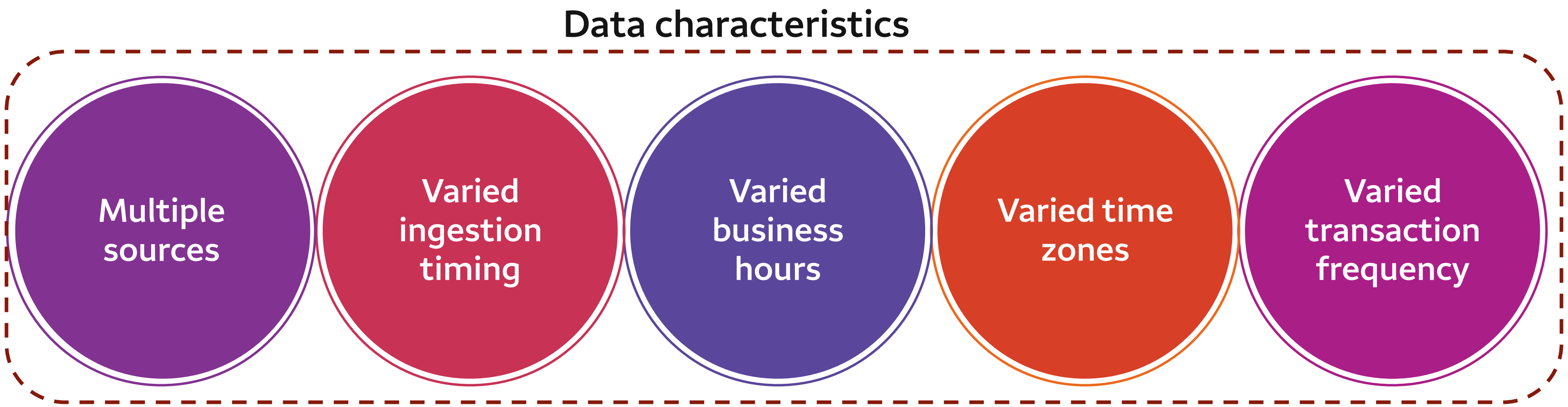
Centralized forecasting can help transform business operations



Typical transaction level data characteristics make multi-level, multi-grain modeling difficult



- Forecasts could be needed every minute
- Each relational input source may have its own reset and refresh time (some at midnight, some early morning, so on)
- Each data source might vary with respect to velocity and volume



Modeling

At what level should one take modeling decisions?

At what level should one take evaluation decisions?

How much approach-standardization should you aim for, given variance and variety in data distribution?

There are a few key decisions when modeling 100s of granular forecasts

.

Business requirement

1. Minute level forecast
2. Daily weekly and monthly forecasts
3. Two pairs of thresholds

** What prerequisites should we satisfy?*

** What should be the basis of selection?*

** What tradeoffs are we dealing with?*

What should I model for?

Approach 1:
Modelling raw volume and value series

Vs.

Approach 2:
Modelling cumulative volume and values series directly

Which algorithm(s) can save the day?

Multiple competing approaches!

- Empirical forecasters (smoothing and exponential family of models)
- Traditional time series (vector autoregressive family, autoregressive moving average family of models)
- Machine learning ensembles
- Deep learning models (prophet, long short-term memory, transformers)

When forecasting 100s of entities, sparsity is always a challenge

Sparsity, in transactions, implies some levels can have nil to poor transaction or usage activity.

Fact of life!
Models revolve around rich data. In its absence, models produce poor results. Rich data ≠ more data

Modeler alert!
Sparsity, by itself, may not be always a bad thing, it could be indicative of patterns!

What are your options?

- Aggregate and disaggregate by time
- Filter inactive segments by time and backfill
- Hypothesis driven features (could run in 1000s)

When it comes to machine learning , there is no one-size fits all approach. Be creative and exhaustive!

A model automation pipeline can help facilitate rapid experimentation

Market offerings like H2O's Driverless AI via its python-client integrated pipelines, use of evolutionary optimization algorithms or Bayesian search methods support model building at scale via multiple automations

- Feature evolution
- Feature transformation
- Target transformation
- Model selection
- Grouped time series
- Modeling documentation
- Model interpretation

Model characteristics

Accuracy

Time

Explainable

Scorer

Number of models, features, and, transformations used in the experiment depends on the dial settings applied on DAI

Model type

Reproducibility

GPU requirement

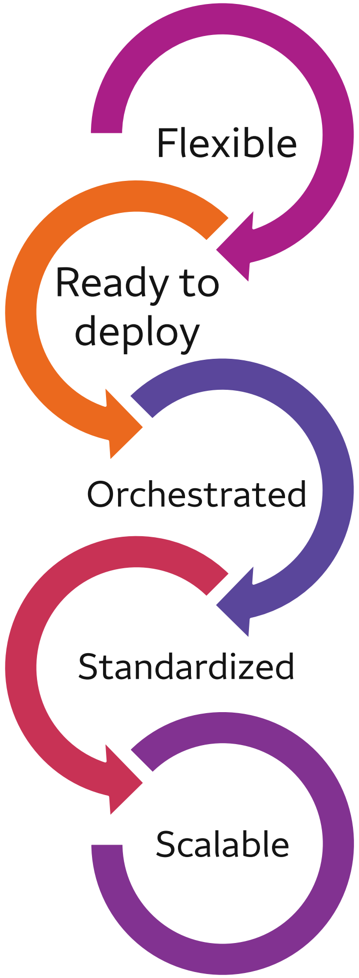
Architecture design

What design principles would be handy when a single technology solution leverages 100s of model objects?

Is it the scale that will primarily drive the direction of solution?

By developing and adapting a reusable and repeatable process, deployment related work becomes streamlined

Deployable model development



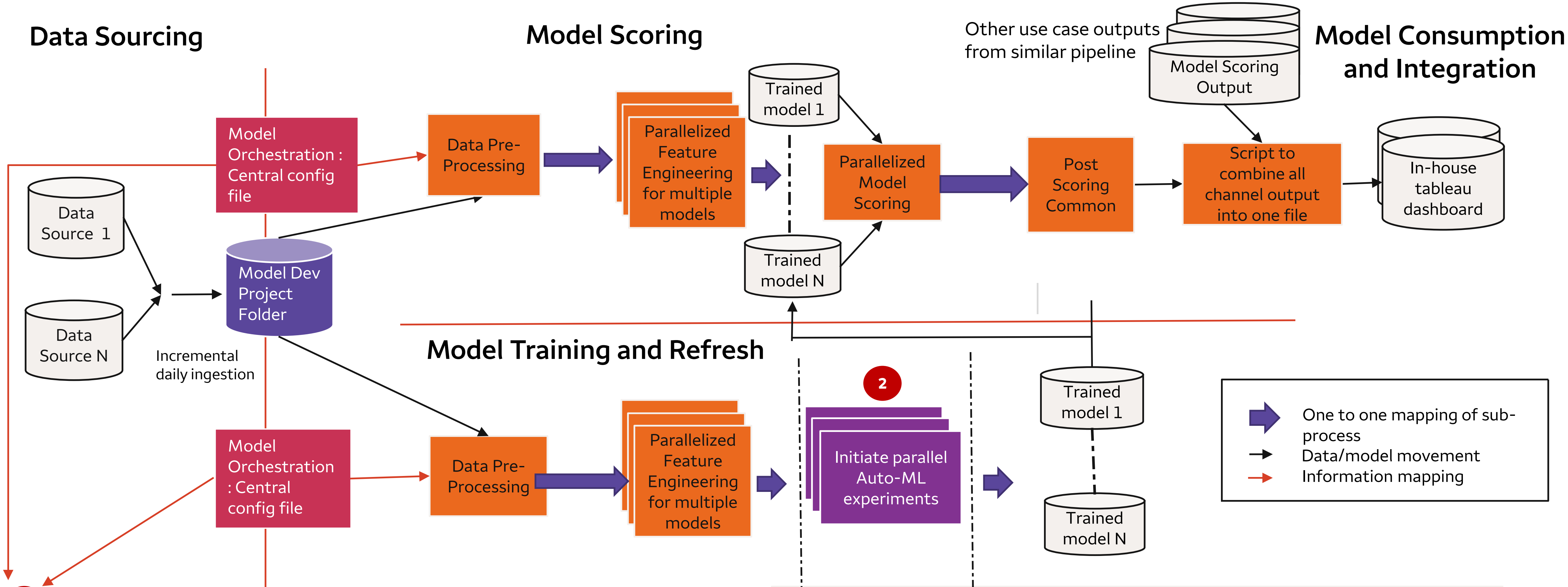
Process entities and characteristics

- Process level utilities
- Process design
- Flow design
- Tools and technological choices
- Integration
- Continuous feedback

Utilities help to organize and maintain re-usable codes organization specific deployment framework

An end state architecture clearly calls out all dependencies and handoffs

A sample architecture template below that shows how we can orchestrate multiple models in one-go



1 **Model orchestration config**– enables model development experimentation – by standardization of model development processes.

Auto-ML component– will accelerate experimentation with fully automated feature selection, model selection and hyper parameter tuning with appropriate cross validation

Deployment

How to maximize efficiency and streamline the last leg?

Deployment requires significant collaboration with technology and clarity on model use

Code hand-off to model operations team

Test end-to-end no intervention flow using flow-orchestration

Framework to generate deployment-ready code for easy hand-off

Pre-deployment

One master code for all channels in one common repository.

Hand-off code base includes multiple channel specific configurations and model objects.

Integration of model output with relevant message streaming service for smooth operations

Deployment decisions

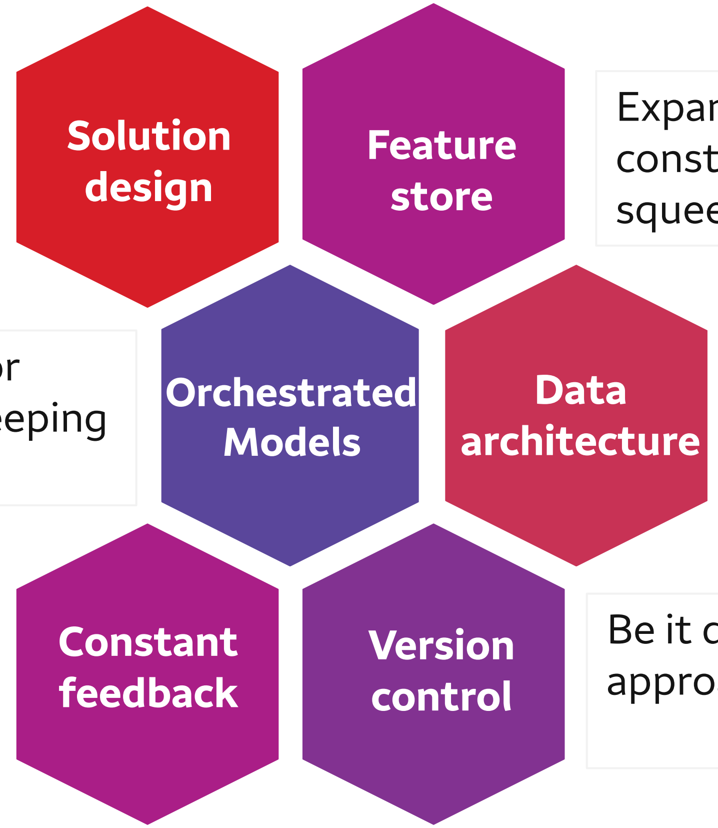
Key guiding principles when creating scalable machine learning solutions for the enterprise

Centrally housed comprehensive knowledge repository of the entire pipeline that can be leveraged by all model developers

Usually rediscovered for Proof of concepts vs. Production and should cover all – data, model, metrics

Expansive at the onset, constantly evolving which squeezes at a model level

Design the pipeline for agility and stability keeping the end state in mind



Matters more than model architecture. Majority of performance bottlenecks here

Test and learn with business partners, subject matter experts, technology

Be it data, pipeline, approach, results

Thank you

Thejas Bhat
Naveen Rathani