BPC - Work Plan for ChemBERTa

Adam Macudzinski, Data Engineering Manager

Executive Summary

Goal: to transform the BPC data architecture to support implementation of the ChemBERTa GNN model

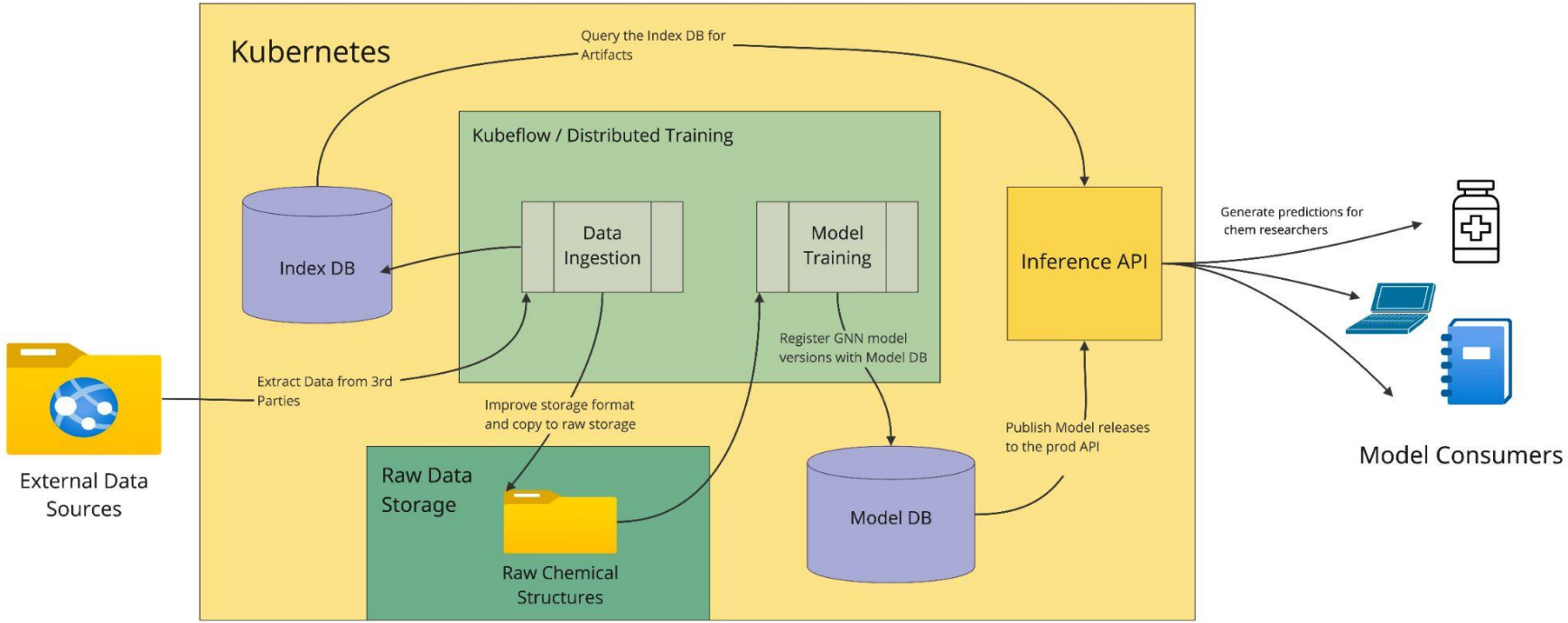
We will focus on 4 key pillars for the system architecture and the initiative planning as well as understand the interdependencies between them to successfully migrate existing data, deploy new architecture, and mitigate risks.

Data Quality and Diversity	Data Storage and Accessibility	Model Performance	Model Deployment
Set data quality benchmarks: the better the data, the better the model. **Expanded data sets**: explore other options beyond full PubChem SMILES.	**Accessible data pipelines**: maximizes modeling agility and minimizes turnaround times **Optimal data structure**: distributed training efficacy strongly depends on the data storage model.	**Performance isn't just accuracy:** beyond accuracy, agility should be optimized for **Robust compute hardware**: cutting edge hardware increases productivity	**Enable iterative improvements**: agile deployments will be key **API deployment**: better API will maximize adoption **Set KPIs**: target usability, runtime performance, reliability, and flexibility

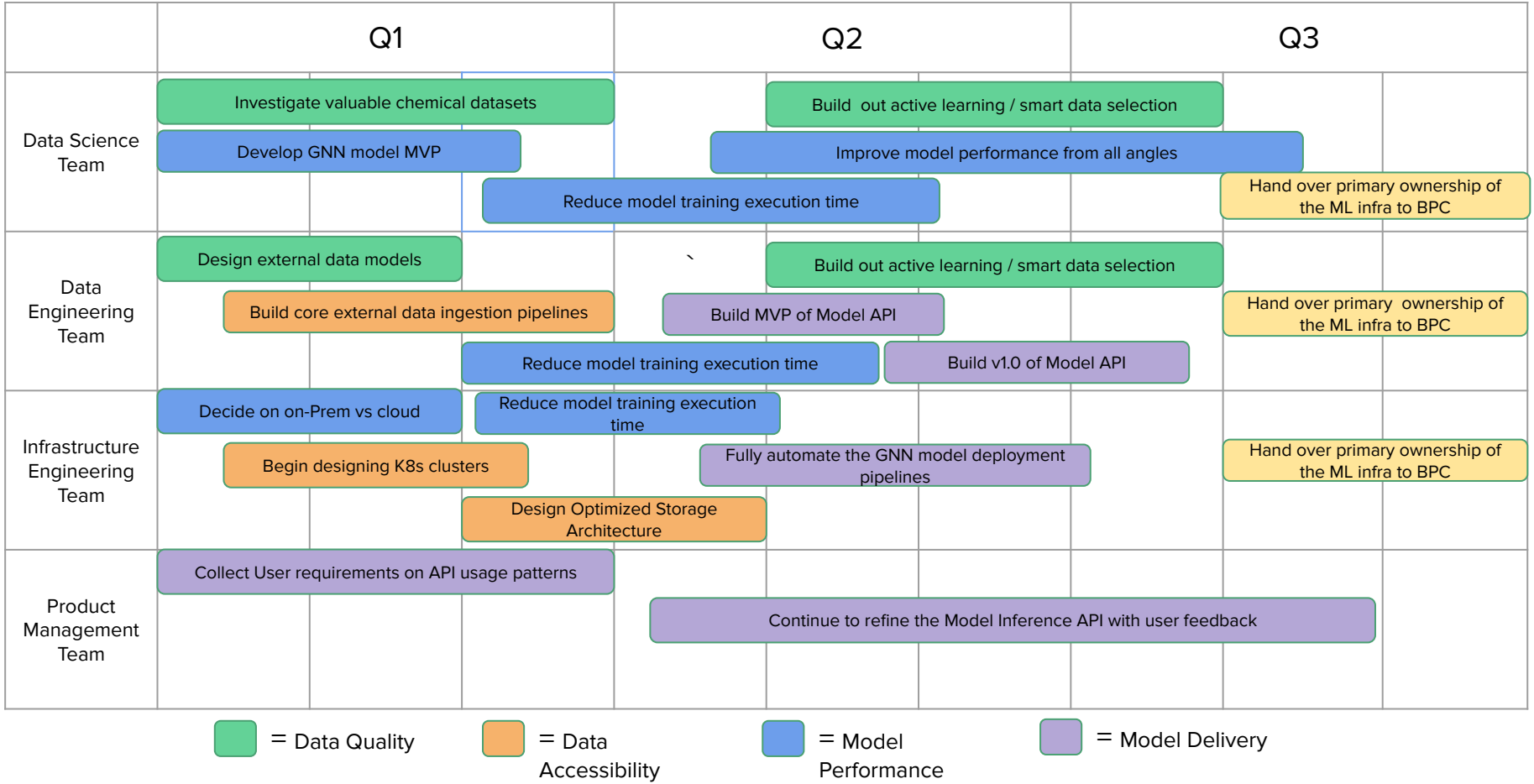
Recommendation: with a team of 8-10 over 3 quarters, we can build a top-of-the-line chemical fingerprint platform that will be able to achieve high velocity molecular property prediction performance.

High-level Software Architecture

Create a strong infrastructure foundation to support machine learning pipeline scalability, data accessibility, and operationalizing of GNN technology for the benefit of cheminformatics.



Project Roadmap



Resourcing recommendation (estimates)

Staffing Resources Recommended:

	Quantity	**Skills Needed**
Data Scientists	3 senior-level	ML, Deep Learning, GNN, Applied Statistics, Tensorflow, Python
Data Engineers	2 senior-level 1 junior-level	Cloud infrastructure, ML Ops, Big Data, Backend, Python, K8s
Infrastructure Engineers	1 senior-level 1 junior-level	DevOps, GitOps, K8s, Linux Admin, NVIDIA, Python
Product Manager	1 technical **Can be BPC internal or Deloitte embedded*	Understanding of API development, and cheminformatics as a nice to have.

Other Costs:

- Compute resources:
  - Cloud: $300-600k