Submitted by: Tagh Sira

Work Plan

Enabling ChemBERTa ML Model API for Chemical Fingerprinting molecular property prediction at Big Pharma Company (BPC)

Executive Summary

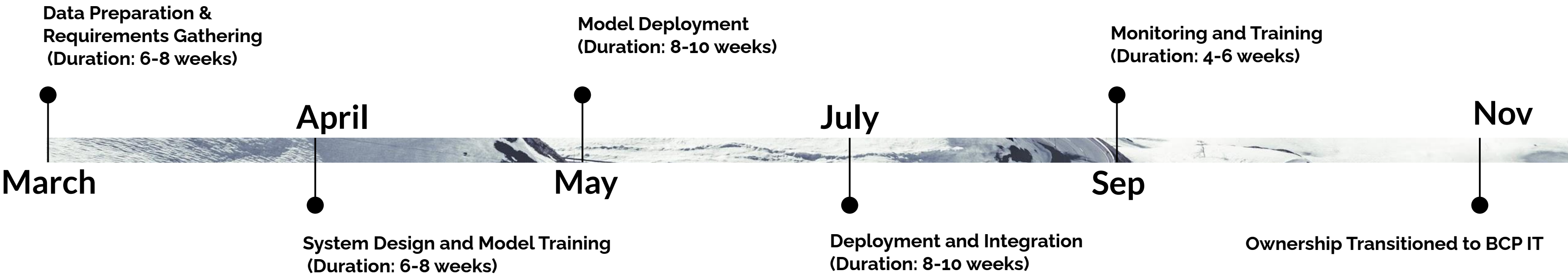
Objective: Develop a solution for chemical fingerprinting to represent molecules for property prediction in downstream applications.

Proposed Approach: Modular approach with 5 phases: Data Preparation, Model Training and Evaluation, Model Deployment, Operations Training, and Continuous Improvement.

Deliverables: Data pipeline, Chemical Fingerprinting Models, Deployment Pipeline, Performance Metrics.

Milestones: Accurate models for chemical fingerprinting, scalable deployment pipeline, continuous improvement mechanism.

Timeline Overview



March

Data Preparation & Requirements Gathering
(Duration: 6-8 weeks)

April

System Design and Model Training
(Duration: 6-8 weeks)

May

Model Deployment
(Duration: 8-10 weeks)

July

Deployment and Integration
(Duration: 8-10 weeks)

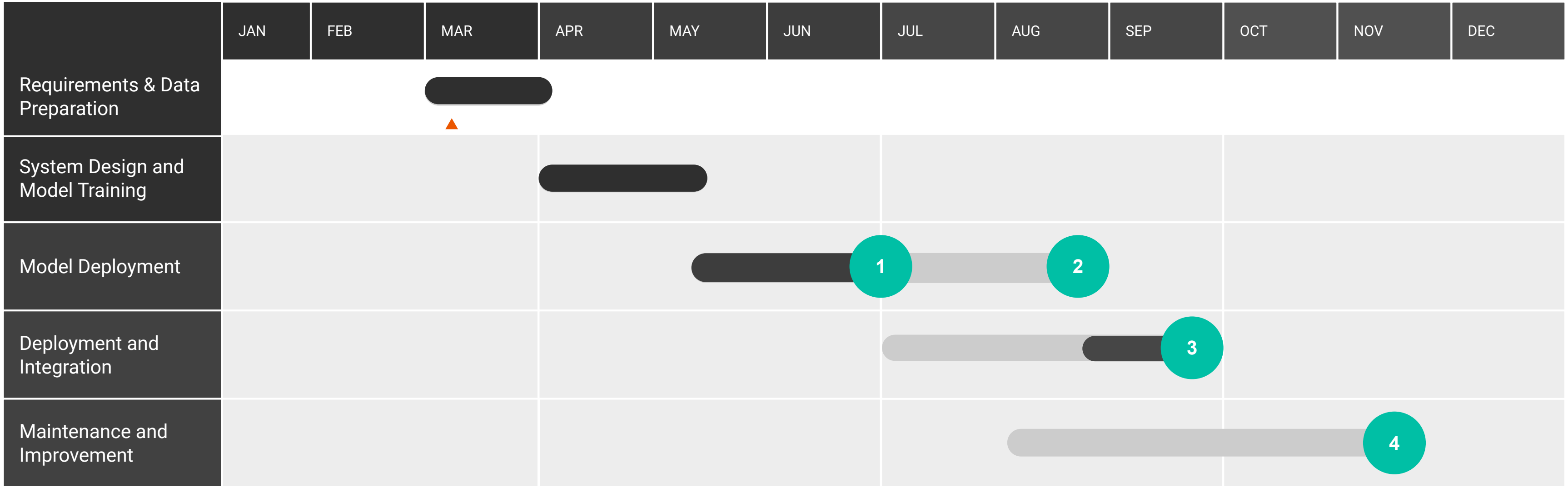
Sep

Monitoring and Training
(Duration: 4-6 weeks)

Nov

Ownership Transitioned to BCP IT

Timeline Phases



JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

Requirements & Data Preparation

System Design and Model Training

Model Deployment

Deployment and Integration

Maintenance and Improvement

Deliverables

1. Chemical Fingerprinting Models
2. Data Pipelines
3. Deployment Pipelines and Observability Monitoring Platform
4. Training, and Documentation provided to BCP ensuring successful adoption and longevity of this solution

Thank you.

Project Value

These are all fictitious values but examples of high level values to present value of this project to BCP

ML Fingerprinting Time Savings

45K

People Hours saved Annually compared to current Manual Fingerprinting Process

Predictive Value

120M

ARR Derived from Predicted Products

Estimated Compute Savings

20K/mo

COGS savings moving off OnPrem compute to Cloud Solution

Required Skill Sets

Data Engineer (Mid to Senior level) - Phase 1-3, 5
Data Scientist (Senior level) - Phase 1-3, 5
DevOps Engineer (Mid to Senior level) - Phase 3, 4
Data Scientist (Senior level) - Phase 1-3, 5
Software Engineer (Mid Level) - Phase 1, 4, 5

Estimated Cost: $650,000 - $900,000 depending on experience and location.

02 | Data Engineering: Data Extraction, Data Transformation, Data Quality, Metadata Management

01 | Data Science: Machine Learning, Statistics, Algorithm Development, Optimization Techniques.

01 | DevOps: Cloud Computing, Infrastructure Design, Scalability, Testing.

02 | Software Engineer (Full Stack): Cloud Platform, Application Engineering for Data Pipeline Integrations.

Phase 1:
Data Preparation

(Duration: 6-8 weeks)

- Understand the existing data sources and format, and evaluate if it can be used to generate chemical fingerprints.
- Develop a data pipeline to extract and transform data into the required format.
- Implement data quality checks to ensure consistency and accuracy of the data.
- Define metadata for the data sources and maintain them.
- Conduct exploratory data analysis to gain insights into the data, including statistical summaries and visualizations.
- Develop a data schema and storage strategy for the processed data.

Phase 2: Model Training and Evaluation

(Duration: 6-8 weeks)

- Develop a set of chemical fingerprinting models based on the specific requirements.
- Generate molecular fingerprints using the selected descriptors, which will be used to represent the molecules in downstream applications.
- Evaluate the effectiveness of the selected descriptors and fingerprints through feature importance analysis and correlation analysis.
- Establish a feedback loop to improve the model performance.

Phase 3: Model Deployment

(Duration: 8-10 weeks)

- Establish a deployment pipeline for the models that enables automation and scalability.
- Deploy the models to the production environment and perform a thorough testing.
- Establish monitoring to ensure the models are performing optimally.
- Evaluate the performance of the models using appropriate metrics and cross-validation techniques.

Phase 4: Deployment and Integration

(Duration: 8-10 weeks)

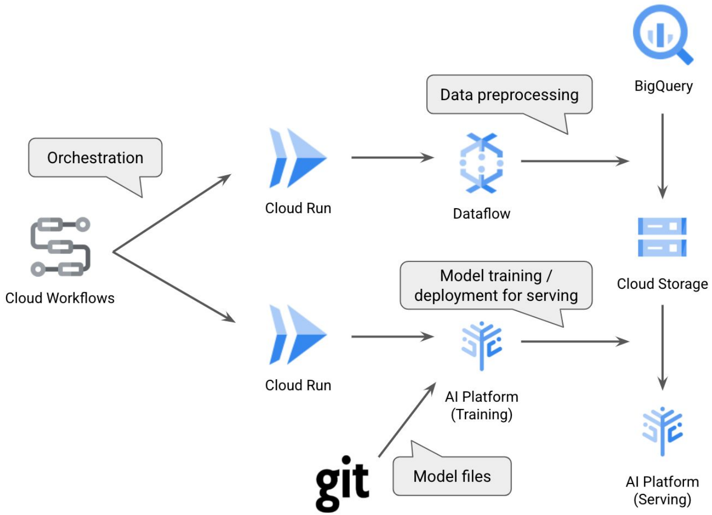
- Develop a mechanism for versioning the models and updating them regularly.
- Deploy the trained models in a production environment, ensuring scalability and performance.
- Integrate the models into downstream applications, such as chemical property prediction or drug discovery.
- Develop a user-friendly interface for interacting with the models and visualizing the results.

Phase 5: Continuous Improvement And Operations Training

(Duration: 4-6 weeks)

- Monitor the performance of the deployed models and collect feedback from users.
- Continuously improve the system through regular updates and maintenance, including data collection and model retraining.
- Provide Training to BCP IT to maintain systems:
  - Update models,
  - Monitor performance,
  - System upgrades,
  - Documentation, Training Resources,
  - Security, Legal, and other Compliance Reviews

Cloud Computing Costs based on Reference Architecture



Operation	Price
Online storage	$0.25 per GB-month (250)
Offline storage	$0.023 per GB-month (23)
Online serving	$0.94 per node per hour (3431)
Batch export	$0.005 per GB (2000)
Streaming ingestion	$0.10 per GB of ingestion (1000)

https://cloud.google.com/products/calculator

Approx Costs: 700 hrs with runtime & 100TB storage = $6700/month