Implementing Advanced AI Capabilities at BPC

October 13, 2024

Executive Summary

Transform BPC's molecular property prediction capabilities by leveraging advanced AI and ML models, specifically implementing ChemBERTa.

Approach

Implement scalable AWS-based cloud infrastructure with GPU capabilities.

Design flexible data pipelines to support modern AI models.

Integrate ChemBERTa model for enhanced molecular property predictions.

Establish future-proof AI workflows for continuous innovation.

Value Proposition

Accelerate drug discovery through improved predictive accuracy.

Enhance operational efficiency with streamlined data processing.

Future-proof research infrastructure for sustained competitive advantage.

Enable seamless integration of emerging AI advancements.

Key Benefits

Improved accuracy with larger datasets.

Faster development of task-specific models.

Adaptable to various molecular representations.

Potential for new chemical insights.

Current Challenges and Need for Transformation

Challenges

Infrastructure lacks GPU capabilities for advanced AI algorithms.

Siloed data systems create inefficiencies in molecular property analysis.

Limited predictive accuracy with traditional models.

Integration difficulties between existing models and new applications.

Need for Transformation

Adopt flexible and scalable AI/ML systems to remain competitive.

Enhance operational efficiency and drive innovation in drug discovery.

Leverage data-driven insights for long-term success.

Position BPC to maintain a leading edge in research capabilities.

Proposed Solution Overview

Scalable Cloud Infrastructure

Implementing an AWS-based cloud infrastructure that provides the necessary GPU capabilities for advanced AI modeling, ensuring scalability as BPC's needs grow.

Flexible Data Pipelines

Designing adaptable data pipelines that can seamlessly integrate various data sources and support modern AI models, enabling efficient data processing and analysis.

Integration of ChemBERTa Model

Utilizing the ChemBERTa model as a foundational tool to enhance molecular property predictions, allowing BPC to leverage cutting-edge AI technologies.

Future AI Workflows

Establishing workflows that can easily incorporate future AI advancements, ensuring BPC remains at the forefront of innovation in drug discovery.

Optional Phase 6 – MLOps Solution Development

Objectives

Establish continuous integration and deployment pipelines to streamline model updates and maintain software quality.

Key Activities

Implement an MLOps framework using AWS tools to automate workflows. Train BPC's team on best practices in MLOps to ensure sustainability.

Continuous Integration

Automate code integration from multiple developers to ensure that changes are tested and integrated regularly.

Continuous Deployment

Ensure that code changes are automatically deployed to production after passing automated tests, minimizing manual intervention.

Monitoring and Feedback

Set up monitoring systems to track model performance in production and gather feedback for improvements.

Future Enhancements

Plan for future MLOps enhancements as the AI landscape evolves, ensuring BPC stays competitive and innovative.

Workflow Enhancements and Future-Proofing

Workflow Enhancements

Implement collaborative tools for improved communication.

Establish standardized APIs for seamless model deployment.

Adopt deep learning frameworks (e.g., PyTorch) for efficient training.

Develop automated ETL processes for data preprocessing.

Future-Proofing Capabilities

Cloud-based architecture allows easy adjustments to workloads.

Ongoing training programs keep team updated on AI advancements.

Flexible infrastructure enables integration of new technologies.

Long-term value through sustainable competitive advantages.

Modular Work Plan and Milestones

Initiation & Planning

ACTIVITIES

Develop and deploy an end-to-end Minimum Viable Product (MVP) system, showcasing the flow from data ingestion to prediction APIs using ChemBERTa.

DELIVERABLES

End-to-end MVP system
Fine-tuned ChemBERTa model
Hands-on demonstration of improved prediction capabilities.

MVP

ACTIVITIES

Establish a secure and scalable AWS environment, including necessary services like EC2 and S3 for data processing and model training.

DELIVERABLES

Configured AWS environment
Secure Virtual Private Cloud (VPC)
Initiated automated data processing pipelines.

Infrastructure

ACTIVITIES

Establish a secure and scalable AWS environment, including necessary services like EC2 and S3 for data processing and model training.

DELIVERABLES

Successful migration of datasets
Data validation and integrity checks
Centralized and scalable data storage.

Full Development

ACTIVITIES

Finalize the development of the ChemBERTa model, ensuring robust integration with downstream applications and optimized performance for predictions.

DELIVERABLES

Fully developed ChemBERTa model
Robust APIs integrated with applications
Enhanced predictive modeling capabilities future-proofing BPC’s ML pipeline.

Training & Handoff

ACTIVITIES

Training sessions and documentation provided to BPC's team.

DELIVERABLES

Empowered teams with knowledge and tools for ongoing use.

MLOps (Optional)

ACTIVITIES

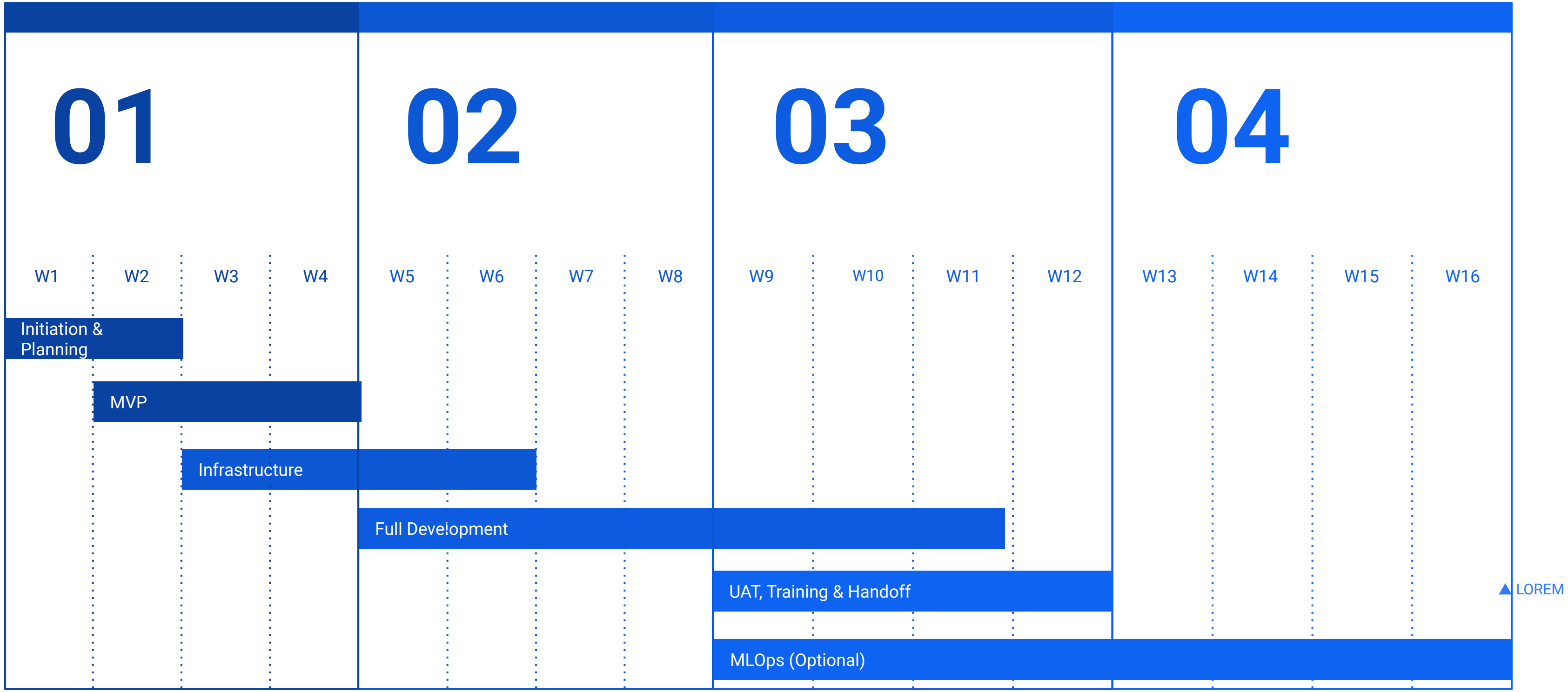
Implemented MLOps framework for continuous model development.

Training and handover of MLOps tools to BPC's team.

DELIVERABLES

Continuous integration and deployment of models.
Ability to rapidly adopt new AI/ML advancements.
Long-term scalability and innovation support.

Roadmap



01

02

03

04

W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13 W14 W15 W16

Initiation & Planning

MVP

Infrastructure

Full Development

UAT, Training & Handoff

MLOps (Optional)

▲ LOREM

Infrastructure Upgrades and Architectural Design

Infrastructure Upgrades

Enhanced Computational Resources

Upgrade to high-performance GPUs for efficient training and inference of ChemBERTa models, enabling the processing of large datasets and complex computations.

Cloud Adoption Strategy

Migrate to AWS to leverage scalable cloud infrastructure that accommodates growth, enhances collaboration, and provides robust data processing capabilities.

Scalable Data Storage Solutions

Implement Amazon S3 for flexible and scalable storage solutions, ensuring efficient handling of large datasets and model artifacts.

Architectural Design

Data Storage Solutions

Utilize Amazon S3 for scalable object storage, separating raw and processed data into dedicated buckets. This allows efficient data retrieval and management, essential for handling large datasets and embeddings.

Compute Resources Planning

Deploy AWS EC2 instances with GPU capabilities (P4d for training, G4dn for inference) to support the intensive computational needs of ChemBERTa. This ensures fast training times and efficient model serving.

Security and Compliance Measures

Implement IAM roles and policies for fine-grained access control, along with data encryption at rest and in transit (using SSL/TLS). Ensure compliance with data protection regulations to safeguard sensitive information.

BPC Capabilities: Before and After Implementation

Before Implementation

- ✗ BPC relied on on-premises storage with limited capacity, resulting in challenges managing large datasets and slow data retrieval times.
- ✗ Machine learning capabilities were constrained to traditional models using engineered features, leading to suboptimal predictive accuracy and inability to handle complex molecular data.
- ✗ Infrastructure was insufficient for deep learning applications, lacking GPU resources and cloud computing, which restricted scalability and efficiency in data processing.

After Implementation

- ✓ Automated data processing pipelines streamline the ingestion and preparation of molecular datasets, reducing time and effort by 40%.
- ✓ ChemBERTa enhances predictive accuracy in property prediction tasks by up to 25% compared to traditional ML models.
- ✓ Scalable cloud infrastructure enables processing of larger datasets, supporting complex modeling tasks without performance bottlenecks.
- ✓ Improved integration of predictive models with downstream applications facilitates real-time predictions and accelerates decision-making processes.

Capabilities with MLOps Phase

After Implementation

- ✓ The MLOps phase allows for automated testing and integration of new models, ensuring that updates can be seamlessly incorporated into existing workflows without disrupting ongoing projects.
- ✓ By employing MLOps practices, BPC can quickly adopt and experiment with new modeling techniques, fostering a culture of innovation that keeps pace with advancements in AI and machine learning.
- ✓ With MLOps, BPC can continuously monitor model performance in production, allowing for timely updates and retraining based on real-world data, thus maintaining predictive accuracy.

Leveraging Expertise and Collaboration

Our Partnership

Joint effort leveraging BPC's domain expertise and our technical capabilities.

Skill Sets and Next Steps

Deloitte:

Data Engineers, ML Engineers, DevOps Specialists

BPC:

Data Experts, IT Personnel, Researchers

Resource Optimization

Phased Resource Allocation for Project Efficiency

Aligning Team Involvement with Project Needs

- Resource allocation is strategically phased to ensure optimal team involvement at each project stage.
- Early phases focus on project management and development expertise to kickstart the initiative.
- Mid-project phases require more data scientists and engineers for model development and integration.
- Final phases emphasize training and support roles to facilitate seamless deployment and knowledge transfer.

Collaboration and Communication Plan

Schedule weekly progress updates to ensure alignment and transparency across teams.

Conduct sprint reviews to evaluate progress, gather feedback, and adjust priorities as needed.

Utilize dedicated communication platforms for efficient information sharing and quick responses.

Establish clear roles and responsibilities for team members to streamline interactions and accountability.

Emphasizing Early UAT Involvement

Ensures Solution Alignment

Involving BPC stakeholders early in User Acceptance Testing (UAT) ensures that the solution aligns with their specific needs and expectations, reducing the risk of misalignment later in the process.

Facilitates Timely Adjustments

Early feedback from BPC experts allows for timely adjustments to the solution, ensuring that any issues are addressed promptly and effectively throughout the development cycle.

Promotes Collaborative Engagement

Active participation in UAT fosters a collaborative environment, encouraging open communication and trust between teams, which enhances overall project success and satisfaction.

Project Cost Breakdown

Training Costs

Training costs total $14,750 for AWS online courses and onsite workshops to enhance team skills.

AWS Infrastructure

AWS infrastructure costs are estimated at $67,788, covering compute and storage services during the project duration.

Personnel Costs

Total personnel costs amount to $405,200, including roles such as project manager, data engineers, and ML engineers.

Additional Costs

Additional costs, including software licenses and a contingency fund, contribute $53,973 to the overall budget.

Total: $541,531

Risk Management

Technical Challenges

Integrating advanced AI models may encounter compatibility issues with existing systems. Continuous testing and validation will ensure smooth integration.

Budget Overruns

Unexpected costs can arise from infrastructure scaling and resource allocation. A contingency fund will be established to manage unforeseen expenses.

Mitigation Strategies

Implementing an early MVP will help validate feasibility and identify potential issues before full-scale deployment.

Next Steps

To move forward, we need to finalize our agreement on the project scope, timelines, and costs.

Next, we will schedule a kick-off meeting to align our teams and set expectations.

This partnership represents a collaborative commitment to innovation and success in advancing BPC's AI capabilities.

Thank You and Contact Information

We Appreciate Your Time and Interest

- We are excited about the potential to collaborate with BPC in implementing advanced AI capabilities.
- Our team is committed to supporting BPC's journey towards enhanced drug discovery and predictive modeling.
- Please feel free to reach out with any questions or further discussions on how we can work together.

Appendix A

Cost Breakdown

3. Cost Tables

3.1 Personnel Costs

Assumptions:

- Project duration: 16 weeks (excluding optional MLOps phase).
- Hourly rates based on industry averages.

Table: Personnel Costs

Role	Quantity	Hourly Rate	Hours/Week	Weeks	Total Hours	Total Cost
Engineering Manager	1	$120	40	16	640	$76,800
Data Engineer	2	$100	40	16	1,280	$128,000
ML Engineer	2	$110	40	16	1,280	$140,800
DevOps Engineer	1	$105	35	16	560	$58,800
Trainer	1	$85	20	4	80	$6,800

| Total Personnel Cost | $411,200 |

3.2 AWS Infrastructure Costs

Assumptions:

- Usage estimates are for the 16-week project duration.

Table: AWS Infrastructure Costs

Service	Estimated Usage	Cost per Unit	Total Cost
Compute Services			
EC2 GPU Instances (Training)	2 instances x 8 hrs/day x 60 days	$24/hour	$23,040
EC2 Instances (General)	4 instances x 8 hrs/day x 80 days	$0.40/hour	$10,240
SageMaker Training Jobs	200 hours of ml.p3.2xlarge	$3.825/hour	$765
SageMaker Endpoints	1 endpoint x 24/7 x 16 weeks	$0.546/hour	$14,707
Storage Services			
Amazon S3 Storage	50 TB/month x 4 months	$0.023/GB/month	$4,600
S3 Data Transfer (Out)	10 TB/month x 4 months	$0.09/GB	$3,600
Data Processing Services			
AWS Glue	200 DPUs x 2 hrs/job x 50 jobs	$0.44/DPU-hour	$8,800
AWS Lambda	2 million requests/month x 4 months	$0.20 per million calls	$1.60
Networking Services			
AWS Direct Connect	Monthly fee + data transfer charges	$1,000/month	$4,000
Management & Monitoring			
Amazon CloudWatch Logs	10 GB/month x 4 months	$0.50/GB ingestion	$20
AWS CloudTrail	Standard (no additional cost)		$0
Other Services			
API Gateway	1 million API calls/month x 4 months	$3.50 per million calls	$14
Miscellaneous AWS Services	Estimated additional costs		$2,000

| Total AWS Infrastructure Cost | $67,787.60 |

3.3 Training Costs

Table: Training Costs

Item	Quantity	Cost per Unit	Total Cost
AWS Online Training Courses	5 users	$500/user	$2,500
Onsite Training Workshops	2 days	$5,000/day	$10,000
Certification Exam Fees	5 exams	$150/exam	$750
Training Materials and Resources			$1,500

| Total Training Cost | | $14,750 |

3.4 Additional Costs and Contingency

Table: Additional Costs

Item	Total Cost
Software Licenses (if applicable)	$5,000
Project Contingency (10% of total)	$48,973
Total Additional Costs	**$53,973**

3.5 Total Estimated Project Cost

Summary Table: Total Project Cost

Category	Total Cost
Personnel Costs	$405,200
AWS Infrastructure Costs	$67,788
Training Costs	$14,750
Additional Costs	$53,973
Grand Total	**$541,711**

3.6 Optional MLOps Phase Costs

Assumptions:

- Additional 8 weeks.
- Involves ML Engineers, DevOps Engineers, and additional AWS costs.

Table: Personnel Costs for MLOps Phase

Role	Quantity	Hourly Rate	Hours/Week	Weeks	Total Hours	Total Cost
ML Engineer	2	$110	40	8	640	$70,400
DevOps Engineer	1	$105	35	8	280	$29,400
Trainer	1	$85	20	2	40	$3,400

| Total Personnel Cost (MLOps Phase) | $103,200 |

Table: AWS Infrastructure Costs for MLOps Phase

Service	Estimated Usage	Cost per Unit	Total Cost
CodePipeline, CodeBuild, etc.	Estimated usage over 8 weeks	$1,000/month	$2,000
Additional EC2 Instances	For MLOps tools	$0.40/hour x 8 weeks	$5,376
Additional S3 Storage	10 TB/month x 2 months	$0.023/GB/month	$460

| Total AWS Infrastructure Cost (MLOps Phase) | $7,836 |

Total Cost for MLOps Phase:

Category	Total Cost
Personnel Costs	$103,200
AWS Infrastructure Costs	$7,836
Total MLOps Phase Cost	**$111,036**

3.7 Updated Grand Total Including MLOps Phase

Grand Total with MLOps Phase:

Category	Total Cost
Initial Project Cost	$541,711
MLOps Phase Cost	$111,036
Updated Grand Total	**$652,747**

Appendix B

Technical Deep Dive: Proposed Solution

High-Level AWS Infrastructure Overview

Virtual Private Cloud (VPC)

The VPC serves as the isolated network environment where all AWS resources are deployed, providing secure communication between components.

Data Ingestion

Data from various sources is ingested through Amazon S3 buckets, where raw data is stored for further processing and analysis.

Compute Resources

Amazon EC2 and SageMaker provide scalable compute resources for model training and inference, utilizing powerful GPU instances for efficient processing.

Model Deployment

Models are deployed using Amazon SageMaker Endpoints, enabling real-time predictions accessible via APIs for downstream applications.

Detailed AWS Network Configuration

VPC Networking

The Virtual Private Cloud (VPC) serves as a secure and isolated network environment for AWS resources. It enables the creation of subnets, route tables, and internet gateways, facilitating controlled access and traffic management.

Public and Private Subnets

Public subnets host resources that require direct internet access, such as load balancers. Private subnets contain resources like EC2 instances that do not require direct internet connectivity, enhancing security.

Route Tables

Route tables define the rules for routing traffic between subnets and to the internet. Each subnet must be associated with a route table, which specifies how to route outbound traffic.

Security Groups

Security groups act as virtual firewalls for instances, controlling inbound and outbound traffic. They allow specific protocols and ports, ensuring that only authorized traffic reaches the AWS resources.

End-to-End Data Flow Process

Data Ingestion

Data is ingested from sources like lab instruments and databases. Files are uploaded to the Raw Data Bucket in S3, triggering event notifications for processing.

Raw data files in S3
S3 event notifications

Data Preprocessing

An AWS Lambda function is triggered by the S3 event to validate data. An AWS Glue ETL job cleans and preprocesses it, storing cleaned data in the Processed Data Bucket.

Validated data
Processed data in S3

Feature Extraction

Using AWS Glue or SageMaker, SMILES strings are tokenized and embeddings generated with ChemBERTa. These embeddings are stored in S3 or the SageMaker Feature Store.

Tokenized sequences
Generated embeddings in S3 or Feature Store

Model Deployment

The trained ChemBERTa model is deployed as a SageMaker Endpoint for real-time inference. An API Gateway exposes the model endpoint for downstream applications.

SageMaker Endpoint
Exposed API for predictions

Detailed Data Processing Pipeline

- Data Upload to S3: Data files (SMILES, SDF, CSV) are uploaded to the S3 Raw Data Bucket, triggering the processing pipeline.

- Data Validation: AWS Lambda functions validate the uploaded files for format and structure, moving invalid files to an error bucket for review.

- Data Cleaning: AWS Glue ETL jobs remove duplicates, handle null entries, and standardize molecular representations to canonical SMILES.

- Data Transformation: Tokenization of SMILES strings occurs, converting them into numerical IDs suitable for model input, adjusting sequences for uniform length.

- Feature Extraction: SageMaker Processing Jobs apply the ChemBERTa model to generate embeddings, which are stored in the S3 Processed Data Bucket or SageMaker Feature Store.

- Data Cataloging: The AWS Glue Data Catalog updates metadata for processed data, ensuring that all transformations are documented for future reference.

Modeling Tech Stack Requirements

- **PyTorch:** A popular deep learning framework used for building and training neural networks, ideal for ChemBERTa's implementation.

- **Hugging Face Transformers:** A library that provides pre-trained transformer models and tools for easy integration and fine-tuning of models like ChemBERTa.

- **AWS SageMaker:** A fully managed service that enables developers to build, train, and deploy machine learning models quickly at scale.

- **NumPy:** A fundamental package for numerical computing in Python, used for handling arrays and mathematical operations essential for data preparation.

- **pandas:** A powerful data manipulation and analysis library that simplifies data handling and preprocessing tasks necessary for model training.

- **scikit-learn:** A library for traditional machine learning algorithms that can complement deep learning approaches in feature engineering and evaluation.

Appendix C

BPC Data Infrastructure and Processing Evaluation

Agenda

- Data Acquisition Methods
- Data Formats Utilized
- Data Acquisition Processes
- Storage Infrastructure
- Database Systems Employed
- Data Backup and Recovery Strategies
- Data Preprocessing Steps
- Feature Extraction Techniques
- Model Training and Evaluation
- Model Deployment Challenges
- Current Computational Resources
- Existing Applications and Tools
- Security and Compliance Measures
- Organizational Structure and Skills
- Identified Gaps in Infrastructure
- Workflow Inefficiencies and Integration Challenges
- Security and Compliance Risks
- Recommendations for Improvement
- Conclusion

Data Acquisition Methods

Internal Research Data

BPC conducts extensive in-house experiments, generating proprietary molecular data from high-throughput screening (HTS), medicinal chemistry efforts, and structure-activity relationship (SAR) studies, amounting to approximately 5 million unique chemical compounds.

Public Databases

BPC supplements its data with publicly available datasets from sources like PubChem, ChEMBL, and DrugBank, integrating around 50 million compounds over the years to enhance its research.

Collaborations and Partnerships

Data is shared through collaborations with academic institutions and biotech firms, governed by confidentiality and licensing agreements to ensure secure and productive partnerships.

Data Formats Utilized

SMILES Strings

SMILES (Simplified Molecular Input Line Entry System) strings are the primary format for representing molecular structures. They allow for efficient storage and computational analysis of chemical structures.

SDF Files

Structure Data Files (SDF) contain detailed structural information along with associated metadata, enabling comprehensive data exchange in cheminformatics tools.

InChI Identifiers

InChI (International Chemical Identifier) provides standardized chemical identifiers that facilitate interoperability and data integration across various systems.

Data Acquisition Processes

Automated Data Capture

Laboratory Information Management Systems (LIMS) streamline the collection of data from laboratory instruments, ensuring real-time data acquisition and reducing manual entry errors.

Manual Data Entry

While necessary for certain data types, manual entry can introduce human error and inconsistencies. Training and guidelines are essential to minimize these risks.

Data Quality Control

Standard Operating Procedures (SOPs) are established to validate and curate data, ensuring accuracy and reliability prior to analysis and usage.

Storage Infrastructure

On-Premises Data Centers

- High-capacity servers equipped with RAID configurations for redundancy.
- Total storage capacity of approximately 500 TB with 70% utilization.
- Environment controlled for optimal performance and data integrity.

Network-Attached Storage (NAS)

- Facilitates shared file storage accessible across various departments.
- Provides centralized data management and quick data retrieval.
- Supports collaborative projects and data sharing among teams.

Database Systems Employed

Oracle Database

BPC primarily utilizes Oracle Database for structured data management, supporting complex queries and robust reporting capabilities. It efficiently handles experimental results and compound metadata, ensuring data integrity and security.

MySQL Instances

Individual departments leverage MySQL instances for specific applications. This relational database management system enables easy access and manipulation of data tailored to departmental needs.

ChemAxon's JChem

ChemAxon's JChem is used for chemical database management, structure searching, and property prediction. It provides advanced cheminformatics capabilities essential for drug discovery processes.

Data Backup and Recovery Strategies

BPC's Data Backup Solutions

- Weekly full backups stored off-site using tape backup technology.
- Daily snapshot backups for critical systems ensure minimal data loss.
- Backup procedures comply with internal data governance and security policies.

Disaster Recovery Plans

- Disaster recovery plan includes a detailed protocol for data restoration.
- Testing of the recovery processes is conducted quarterly to ensure effectiveness.
- Limitations exist in recovery times for mission-critical applications, affecting operational continuity.

Data Preprocessing Steps

Data Cleaning Techniques

BPC employs tools like Open Babel and custom in-house scripts to remove duplicates and inconsistencies in the dataset, ensuring a reliable and accurate data foundation.

Standardization Methods

Canonical SMILES conversion is utilized for ensuring uniform molecular representations, while normalization processes adjust factors like pH and stereochemistry for consistency.

Validation and Quality Assurance

Data validation protocols are implemented as part of the cleaning process, focusing on maintaining high-quality data through established standard operating procedures (SOPs).

Feature Extraction Techniques

Chemical Fingerprinting

BPC employs various chemical fingerprinting techniques such as Morgan (circular) fingerprints, which generate 1024-bit vectors, and MACCS keys, comprising 166-bit structural keys, to represent molecular structures.

Descriptors Calculation

Physicochemical properties like molecular weight, logP, and hydrogen bond donors/acceptors are calculated to provide essential features for predictive modeling and analysis.

Software Tools Utilized

Tools like RDKit are pivotal for generating fingerprints and calculating descriptors, while KNIME Analytics Platform aids in visual workflow for data processing and feature extraction.

Model Training and Evaluation

Machine Learning Models

BPC employs several machine learning algorithms, including Random Forests, Support Vector Machines (SVMs), and Gradient Boosting Machines, to predict biological activity and ADMET properties.

Computational Resources

The modeling process utilizes high-performance local workstations equipped with multi-core CPUs and 32-64 GB of RAM, alongside a small compute cluster of 10 nodes for model training.

Target Properties

Models are trained to predict key properties such as biological activity, absorption, distribution, metabolism, excretion, and toxicity (ADMET), which are crucial for drug discovery.

Model Deployment Challenges

Manual Deployment Limitations

- Manual deployment processes are time-consuming, requiring extensive human intervention.
- Higher likelihood of errors during deployment leads to inconsistent model performance.
- Limited scalability and adaptability to changing requirements or data inputs.

Integration Issues with Applications

- Ad-hoc integrations create challenges in maintaining model compatibility with applications.
- Frequent need for custom scripts increases technical debt and maintenance costs.
- Difficulties in monitoring model performance post-deployment due to lack of integration tools.

Current Computational Resources

Hardware Specifications

- Multi-core CPUs: Intel Xeon processors with up to 16 cores per server.
- Limited GPU availability: A few NVIDIA GTX 1080 cards used for testing purposes.
- Storage systems: Combination of HDDs for bulk storage and SSDs for high-performance tasks.
- On-premises servers: Dedicated servers with RAID configurations for data redundancy.

Software and Tools

- Operating Systems: CentOS and Ubuntu for Linux servers; Windows for workstations.
- Primary Programming Languages: Python for data processing; R for statistical analysis.
- Machine Learning Libraries: scikit-learn, pandas, NumPy for various data tasks.
- Cheminformatics Tools: RDKit and ChemAxon software for chemical data processing.

Existing Applications and Tools

01

BPC employs a custom in-house drug discovery platform for managing compound libraries and conducting virtual screenings.

02

The Schrödinger Suite is utilized for molecular modeling and advanced simulations in drug discovery.

03

MOE (Molecular Operating Environment) serves as a comprehensive software suite for both computational and cheminformatics tasks.

04

Spotfire is used for data visualization, allowing researchers to explore complex datasets interactively.

05

Tableau is leveraged for creating dashboards and visual analytics, enhancing decision-making processes.

06

Jupyter Notebooks are employed by data scientists for exploratory data analysis and sharing insights collaboratively.

Security and Compliance Measures

Data Security Practices at BPC

- Access controls are implemented through role-based access management to safeguard sensitive data.
- Encryption is applied to data at rest and in transit, using SSL/TLS protocols for secure communications.
- Regular security audits and vulnerability assessments are conducted to identify potential risks.

Regulatory Compliance Framework

- BPC adheres to Good Laboratory Practice (GLP) guidelines to ensure quality in laboratory operations.
- Compliance with GDPR and HIPAA is monitored, particularly for datasets containing personal information.
- Policies are in place for data retention and disposal to meet compliance requirements.

Organizational Structure and Skills

IT Department Overview

The IT department comprises 20 personnel responsible for infrastructure, networking, and technical support. They possess strong skills in traditional IT management but lack specialized experience in cloud solutions.

Data Science Team Composition

The data science team includes 5 data scientists with backgrounds in chemistry and basic machine learning. Their expertise lies in traditional statistical methods, but they have limited exposure to advanced AI/ML techniques.

Identified Skills Gaps

Key gaps include insufficient knowledge in AI/ML frameworks, particularly in deep learning and transformer models, along with a lack of experience in cloud computing services and MLOps practices.

Identified Gaps in Infrastructure

Computational Resource Limitations

- Insufficient availability of GPUs for deep learning tasks.
- Limited high-performance computing (HPC) resources hinder complex model training.
- Current compute cluster lacks capability for large-scale AI/ML workloads.

Storage System Constraints

- Existing storage solutions are nearing capacity and lack scalability.
- Inability to handle increased data volume from advanced modeling techniques.
- Current infrastructure not designed for efficient data retrieval and management.

Workflow Inefficiencies and Integration Challenges

Manual Processes

BPC heavily relies on manual data entry and management, leading to increased error rates and inefficiencies in data processing.

Fragmented Data Ecosystem

Data is stored across multiple systems without a centralized platform, resulting in difficulties in accessing and integrating information.

Data Format Incompatibilities

Inconsistent data formats require frequent conversions, which can introduce errors and delay analysis, hindering timely decision-making.

Security and Compliance Risks

01

Inconsistent application of access controls increases the risk of unauthorized data access across systems.

02

Limited data encryption practices leave sensitive information vulnerable to breaches and cyber threats.

03

Non-compliance with regulations such as GDPR and HIPAA can lead to legal penalties and reputational damage for the organization.

Recommendations for Improvement

Enhance Computational Infrastructure

Invest in high-performance computing resources, including GPUs and cloud solutions, to support AI/ML workloads efficiently.

Automate Workflow Processes

Implement workflow management tools like Apache Airflow to automate data processing and integration, reducing manual intervention.

Upskill Workforce

Provide comprehensive training programs in AI/ML, cloud computing, and data governance to bridge the skills gap within teams.

Standardize Data Formats

Establish organization-wide data governance policies to ensure consistent data formats and facilitate seamless integration across systems.

Strengthen Security Measures

Implement robust security protocols including end-to-end encryption and strict access controls to protect sensitive data.

Integrate Compliance Frameworks

Embed regulatory compliance checks into IT processes to ensure adherence to GxP, GDPR, and HIPAA requirements across all data operations.

Conclusion

The audit of BPC's data infrastructure and processing capabilities reveals significant strengths, particularly in the volume of data collected and existing applications. However, crucial gaps in computational resources, workflow inefficiencies, and integration challenges have been identified. Addressing these gaps is essential for enhancing data management capabilities and optimizing the implementation of advanced models like ChemBERTa. By investing in infrastructure upgrades, automating workflows, and upskilling staff, BPC can align its operations with its strategic goals and improve overall efficiency.

END