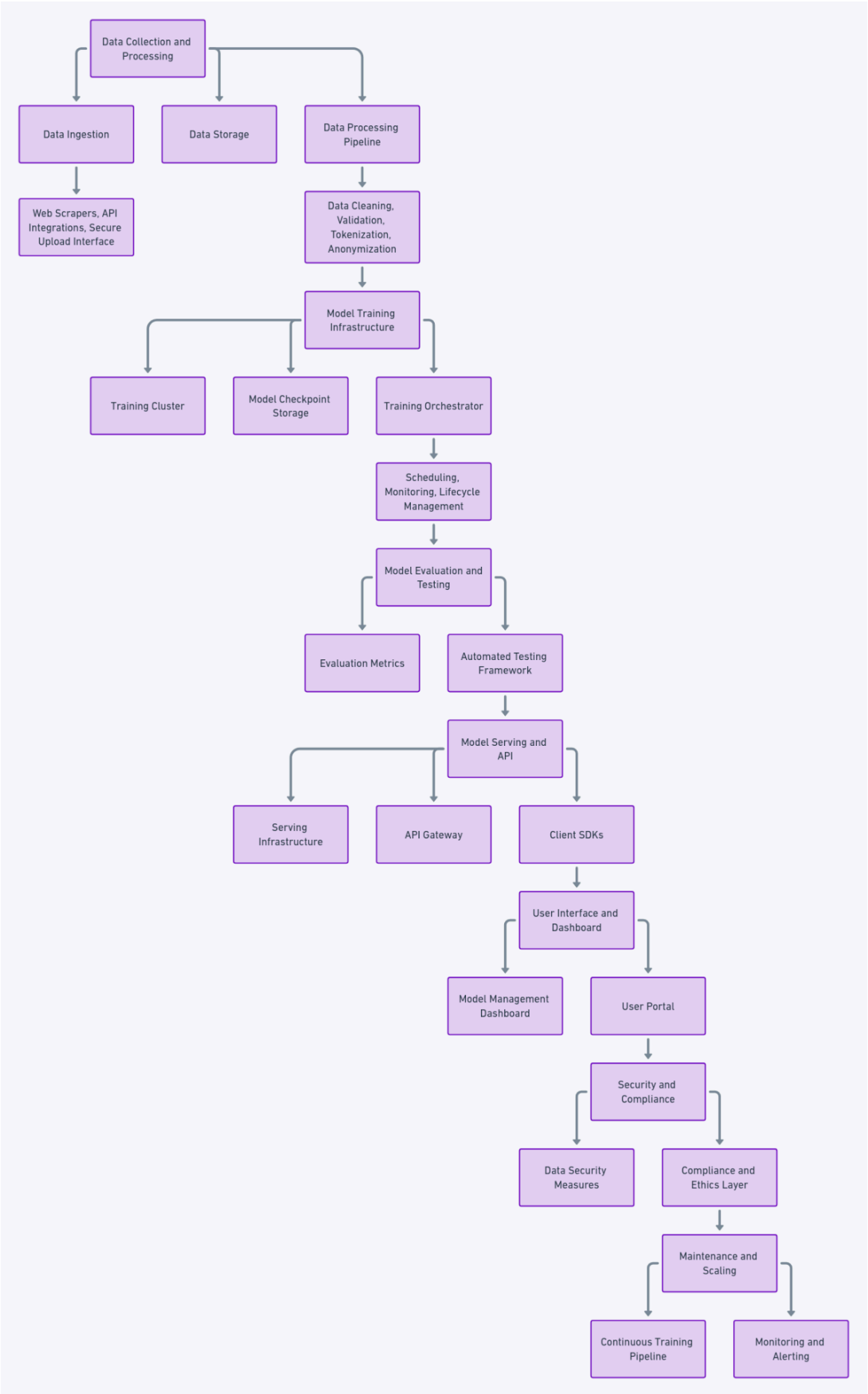
Athena Architecture Design Documentation

Data Collection and Processing

- **Data Ingestion:** The system will feature mechanisms for pulling data from diverse sources, including web scrapers to collect public data, API integrations to gather structured data, and a secure user interface for manual data uploads. Data provenance will be tracked to ensure traceability and compliance with legal requirements.
- **Data Storage:** A scalable and secure data storage solution will be implemented, capable of handling petabytes of data with redundancy and backup mechanisms in place. It will support fast read/write operations required during the training phase.
- **Data Processing Pipeline:** The processing pipeline will be responsible for standardizing and preparing data for training. This includes noise reduction, tokenization, and anonymization to protect privacy. The pipeline will be modular to easily adapt to new data types and sources.

Architecture design figma link :
https://www.figma.com/file/aXGiEjkOxbd22sTwwq8Qhr/Athena-architecture?type=whiteboard&node-id=0-1&t=B6fJJwwgSvLpf0iC-0


Data Collection and Processing
Data Ingestion
Data Storage
Data Processing Pipeline
Web Scrapers, API Integrations, Secure Upload Interface
Data Cleaning, Validation, Tokenization, Anonymization
Model Training Infrastructure
Training Cluster
Model Checkpoint Storage
Training Orchestrator
Scheduling, Monitoring, Lifecycle Management
Model Evaluation and Testing
Evaluation Metrics
Automated Testing Framework
Model Serving and API
Serving Infrastructure
API Gateway
Client SDKs
User Interface and Dashboard
Model Management Dashboard
User Portal
Security and Compliance
Data Security Measures
Compliance and Ethics Layer
Maintenance and Scaling
Continuous Training Pipeline
Monitoring and Alerting

Model Training Infrastructure

- **Training Cluster:** A set of high-performance GPU or TPU servers, likely provisioned in the cloud for scalability. They will be equipped with the necessary frameworks and libraries for distributed deep learning.
- **Model Checkpoint Storage:** Model states will be periodically saved during training to allow for recovery in case of interruptions. This storage will be optimized for high-speed read/write operations to minimize downtime.
- **Training Orchestrator:** This component will manage the orchestration of training jobs, monitoring resource utilization, scheduling jobs based on resource availability, and handling failover scenarios. It will also enable fine-grained tracking of model versions and lineage.

Model Evaluation and Testing

- **Evaluation Metrics:** The model will be evaluated against a set of predefined metrics that may include perplexity, BLEU score for translation tasks, F1 score for classification tasks, and fairness metrics to ensure that the model does not exhibit bias.
- **Automated Testing Framework:** A suite of automated tests will ensure that every update to the model maintains or improves the performance standards. It will also perform load testing to simulate real-world use cases.

Model Serving and API

- **Serving Infrastructure:** The model will be served using a scalable infrastructure that supports containerization for easy deployment and auto-scaling to handle varying load levels.
- **API Gateway:** A secure API gateway will manage external access to the model, providing features like authentication, rate limiting, and logging of requests for auditing and optimization purposes.
- **Client SDKs:** SDKs for various programming languages will make it easier for developers to integrate Athena into applications, providing simplified interfaces for common tasks and documentation on best practices.

User Interface and Dashboard

- **Model Management Dashboard:** Administrators will use this dashboard to monitor the model's performance, manage deployments, and view logs and metrics. It will facilitate a clear overview of the model's health and usage patterns.
- **User Portal:** A user-friendly interface for contributors to submit data, manage their contributions, and interact with the model, such as generating predictions or accessing model insights.

Security and Compliance

- **Data Security Measures:** All data will be encrypted in transit and at rest. Access controls will be implemented using the principle of least privilege, and regular security audits will be conducted.
- **Compliance and Ethics Layer:** This will ensure the model adheres to ethical guidelines and compliance standards. It will include tools for bias detection and mitigation, and reporting mechanisms for compliance with regulations like GDPR.

Maintenance and Scaling

- **Continuous Training Pipeline:** To keep Athena current and improving, there will be a pipeline for continuous training that allows the model to learn from new data without service interruptions.
- **Monitoring and Alerting:** Robust monitoring for operational metrics and model performance, with alert systems in place for any anomalies or degradation, ensuring high availability and reliability.

Data Source

- **Common Crawl:** This is a massive and diverse dataset comprising terabytes of raw web data from billions of web pages, updated monthly.
- **RefinedWeb:** It's a curated dataset with more than 5 trillion tokens of textual data, with 600 billion tokens made publicly available. This data is deduplicated and filtered from Common Crawl and is particularly aimed at training models like Falcon-40B.

- **The Pile**: An 800 GB corpus created from 22 diverse datasets, primarily from academic or professional sources, which is designed to improve a model's generalization capability.
- **C4 (Colossal Clean Crawled Corpus):** A 750 GB English corpus derived from Common Crawl with heavy deduplication and cleaning to remove non-natural language data.
- **Starcoder Data:** A programming-centric dataset built from GitHub and Jupyter Notebooks, containing 250 billion tokens across 86 programming languages.
- BookCorpus: A dataset made from scraped data of 11,000 unpublished books, totaling about 985 million words.
- **ROOTS:** A multilingual dataset of 1.6TB curated from texts in 59 languages, heavily deduplicated and filtered from Common Crawl and other sources.

Wikipedia Dataset: This dataset contains cleaned text from Wikipedia in all available languages. The English Wikipedia dataset alone provides nearly 20 GB of data.