1. Folder Structure:

Raw Zone:

This zone stores raw data directly from the source, without any processing.

- `/raw/clinical_data/`
  - `/appointments/YY/MM/DD` - Data from appointment systems (e.g., scheduling, visit notes).
  - `/patient_records/YY/MM/DD` - Structured data such as electronic health records (EHR), lab results.
- `/raw/imaging_data/`
  - `/x-rays/YY/MM/DD` - X-ray images and metadata.
  - `/ct_scans/YY/MM/DD` - CT scans and associated metadata.
  - `/mri_scans/YY/MM/DD` - MRI scans and metadata.
- `/raw/genomic_data/`
  - `/sequencing_data/YY/MM/DD` - Raw genomic sequences from next-generation sequencing (NGS).
  - `/variant_data/YY/MM/DD` - Variants and annotations from genomic analysis.
- `/raw/research_studies/`
  - `/clinical_trials/YY/MM/DD` - Data from ongoing and completed clinical trials.
  - `/patient_surveys/YY/MM/DD` - Survey data collected for research purposes.

Cleansed Zone:

This zone stores processed data that has been cleaned and standardized.

- `/cleansed/clinical_data/`
  - `/standardized_patient_records/YY/MM/DD` - Standardized EHR and patient records.
  - `/aggregated_lab_results/YY/MM/DD` - Aggregated and cleaned lab results.
- `/cleansed/imaging_data/`
  - `/formatted_xrays/YY/MM/DD` - X-ray images converted into standard formats (e.g., DICOM).
  - `/processed_ct_scans/YY/MM/DD` - Processed and indexed CT scans.
  - `/segmented_mri_scans/` - Segmented MRI scans for analysis.
- `/cleansed/genomic_data/`
  - `/aligned_sequences/YY/MM/DD` - Aligned genomic sequences for further analysis.
  - `/annotated_variants/YY/MM/DD` - Cleaned and annotated genomic variants.
- `/cleansed/research_studies/`
  - `/cleaned_clinical_trials/YY/MM/DD` - Cleaned data from clinical trials.
  - `/standardized_surveys/YY/MM/DD` - Standardized patient survey data.

Curated Zone:

The curated zone contains data that has been integrated and enriched, ready for analysis and reporting.

- **Clinical Data:**
  - `/curated/patient_profiles/` - Integrated patient profiles combining EHR, lab results, and appointment data.
  - `/curated/disease_outcomes/` - Aggregated data on disease outcomes and treatment effectiveness.
- **Imaging Data:**
  - `/curated/imaging_analysis/`- Results of imaging analysis, such as tumor detection or segmentation.
  - `/curated/diagnostic_imaging/` - Curated and annotated imaging data used for diagnostic purposes.
- **Genomic Data:**
  - `/curated/genomic_variants/` - Curated variant data for research or clinical use.
  - `/curated/genomic_profiles/` - Integrated genomic profiles for patients.
- **Research Data:**
  - `/curated/clinical_trial_results/` - Consolidated results from clinical trials.
  - `/curated/research_findings/` - Key findings from research studies.

Analytics Zone:

This zone contains data processed specifically for analytical tasks, such as machine learning models or reporting.

- **Predictive Analytics:**
  - `/analytics/patient_risk_predictions/` - Predictive models identifying at-risk patients.
  - `/analytics/disease_progression/` - Models predicting disease progression.
- **Imaging Analytics:**
  - `/analytics/diagnostic_models/` - Machine learning models applied to imaging data.
  - `/analytics/segmentation_results/` - Results from image segmentation models.
- **Genomic Analytics:**
  - `/analytics/genomic_risk_factors/` - Analysis of genomic data to identify risk factors for diseases.
  - `/analytics/precision_medicine/` - Models for personalized treatment based on genomic data.

2. Data Governance:

Effective data governance is crucial for maintaining data quality, privacy, and interoperability across healthcare data sources.

- **Privacy and Compliance:**
  - **HIPAA Compliance:** Ensure that all patient data is handled in accordance with the Health Insurance Portability and Accountability Act (HIPAA). This includes encryption (both at rest and in transit) and strict access controls.

    - **Data Masking:** Apply data masking techniques to protect sensitive information such as patient identifiers during analysis and reporting.
- **Access Control and Auditing:**
    - Implement Role-Based Access Control (RBAC) to ensure that only authorized personnel have access to sensitive data.
    - Enable detailed logging and auditing to track access and modifications to data.

3. Data Processing Pipelines:

To handle both structured and unstructured data, the Data Lake should have robust data processing pipelines that can scale efficiently.

- **Structured Data (e.g., Patient Records):**
    - **Ingestion:** Use ETL/ELT tools like Apache NiFi or Azure Data Factory to ingest structured data from appointment systems and EHR.
    - **Processing:** Utilize Apache Spark or Azure Synapse for data cleansing, deduplication, and transformation.
    - **Storage:** Store structured data in the Cleansed and Curated zones for further analysis.
- **Unstructured Data (e.g., Medical Images):**
    - **Ingestion:** Use specialized ingestion pipelines for large, unstructured data such as DICOM files (e.g., using tools like Apache Kafka or AWS S3 for real-time ingestion).
    - **Processing:** Apply image processing techniques using frameworks like TensorFlow or PyTorch. This could include tasks such as image segmentation, classification, and object detection.
    - **Storage:** Store the processed images and metadata in the Cleansed and Curated zones, ensuring that they are indexed and searchable.
- **Genomic Data:**
    - **Ingestion:** Utilize pipelines specifically designed for high-throughput sequencing data (e.g., Apache Parquet for efficient storage of large genomic datasets).
    - **Processing:** Perform sequence alignment, variant calling, and annotation using tools like GATK or SAMtools.
    - **Storage:** Store the processed genomic data in a format that allows for efficient querying and analysis in the Cleansed and Curated zones.

Workflow Example:

1. **Ingestion:** Patient records and appointment data are ingested into the `/raw/clinical_data/` directory. At the same time, medical images from MRI scans are ingested into `/raw/imaging_data/`.
2. **Processing:** The raw patient records are cleaned and standardized in the `/cleansed/clinical_data/standardized_patient_records/` directory, while MRI scans are processed and indexed in `/cleansed/imaging_data/processed_mri_scans/`.
3. **Curation:** The processed clinical data is integrated with genomic data to create a comprehensive patient profile in `/curated/patient_profiles/`.
4. **Analysis:** Machine learning models are run on the curated data to predict patient risk, with results stored in `/analytics/patient_risk_predictions/`.

5. **Governance:** Data access is controlled via RBAC, and compliance with HIPAA is ensured through encryption and regular audits.