Proposed Cloud-Based Data Flow for HE^{2}AT Center Post-Funding Restructure

Context and Purpose

Due to NIH's withdrawal of funding for South African partners, the HE^{2}AT Center's data infrastructure must transition to a cloud-based solution (e.g., Azure). This document outlines a proposed data flow diagram indicating:

- Levels of access for different HE^{2}AT partners
- Data ownership and control boundaries
- Data storage policies compliant with POPIA and GDPR
- Role of CeSHHAR and data access restrictions (e.g., date of birth)
- Long-term archival and use of de-identified data for future research
- Data security through geographic masking techniques

Data Access Roles and Responsibilities

- **Data Providers**: Retain ownership of Original Study Data. Data is stored in-region via secure transfer protocols and only accessed by the Core Data Team.
- **Core Data Team (UCT)**: Handles pre-processing, harmonisation, validation, and de-identification. Maintains audit logs, transformation records, and ensures metadata fidelity. Applies jittering and geographic aggregation methods.
- **Azure Cloud Platform**: Secure, compliant repository hosting geographically partitioned storage accounts with scalable role-based access control (RBAC).
- **HE^{2}AT Consortium (CeSHHAR, IBM, WHC, UPGC)**: Access to Consortium Shared Data, controlled via RBAC. No direct access to identifiable information.
- **External Researchers**: Access only fully de-identified datasets following DAC review, DTA signing, and full compliance audits.

Azure Cloud Technical Architecture

- **Geographically Scoped Storage**: Azure Storage Accounts are region-specific (e.g., South Africa North, West Europe) to comply with POPIA and GDPR.
- **Data Containers**: Containers are separated for raw, harmonised, and de-identified data by project, access level, and jurisdiction.
- **Access Levels**:
  - Level 0: Original Study Data (Core Data Team only)

  - Level 1: Consortium Shared Data (HE^{2}AT Consortium via RBAC)
  - Level 2: RP1/RP2 De-identified Data (DAC-approved)
  - Level 3: Inferential Data (open, non-identifiable aggregates)

- **Access Management**: Azure Active Directory manages RBAC tiers. Conditional access restricts login by IP, geolocation, and 2FA.
- **Encryption and Compliance**: AES-256 encryption, TLS in transit. Azure Key Vault secures key lifecycle. Meets NIST and ISO standards.
- **Monitoring and Auditing**: Continuous tracking via Azure Monitor, Log Analytics, and Sentinel with incident alerts.

Data Security and Jittering Techniques

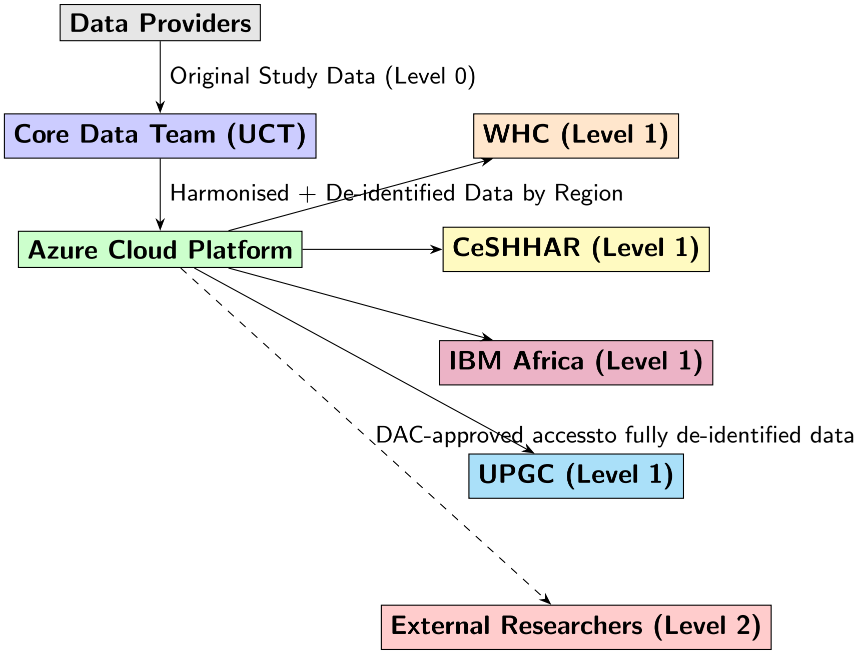
- **Geographic Aggregation**: Data aggregated to census small areas or wards to reduce re-identification risk.
- **Location Jittering**: Gaussian displacement used to obfuscate coordinates, respecting local population density and spatial k-anonymity.
- **Expert Review**: Geo-masking approaches reviewed by a technical committee from UCT, IBM, and NIH.

Data Use and DTA Structure

- **Data Use**: De-identified datasets may be reused for future research projects under ethical clearance and DAC oversight.
- **DTA Framework**: All DTAs will specify:
  - Dataset scope and project affiliation
  - Region of storage and compliance standards
  - Permitted duration and modality of access
  - Non-transfer clauses and audit rights

Data Flow Diagram Overview

The diagram below will be further translated into XML/Draw.IO for presentation.


Data Providers
Original Study Data (Level 0)
Core Data Team (UCT)
Harmonised + De-identified Data by Region
Azure Cloud Platform
WHC (Level 1)
CeSHHAR (Level 1)
IBM Africa (Level 1)
UPGC (Level 1)
DAC-approved accessto fully de-identified data
External Researchers (Level 2)

Key Notes for Data Providers

- Data ownership remains with the original provider.
- DTAs ensure transparent roles, rights, and revocability.
- Regional isolation prevents unlawful cross-border data transfer.
- Data is stored, accessed, and used under secure, compliant, and ethically reviewed protocols.
- Extended future use of de-identified datasets is possible, always under DAC oversight and updated ethical approval.