[Open]

Technical / Operational Paperwork:

Technical & Operational Implementation Plan

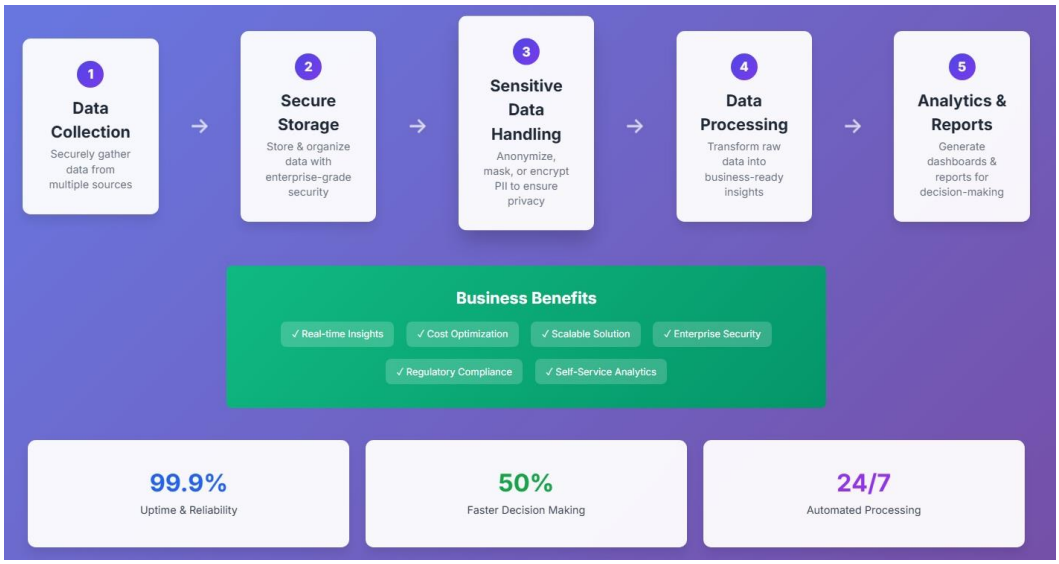
Prepared For: Internal Data/IT Teams, Consultants & Implementation Partners

Date: July 4, 2025

1. Technical Architecture Overview

1.1 System Architecture Philosophy

This platform is built on a cloud-native, microservices architecture emphasizing scalability, security, and maintainability. It adheres to industry best practices, including event-driven processing, API-first design, and zero-trust security principles.

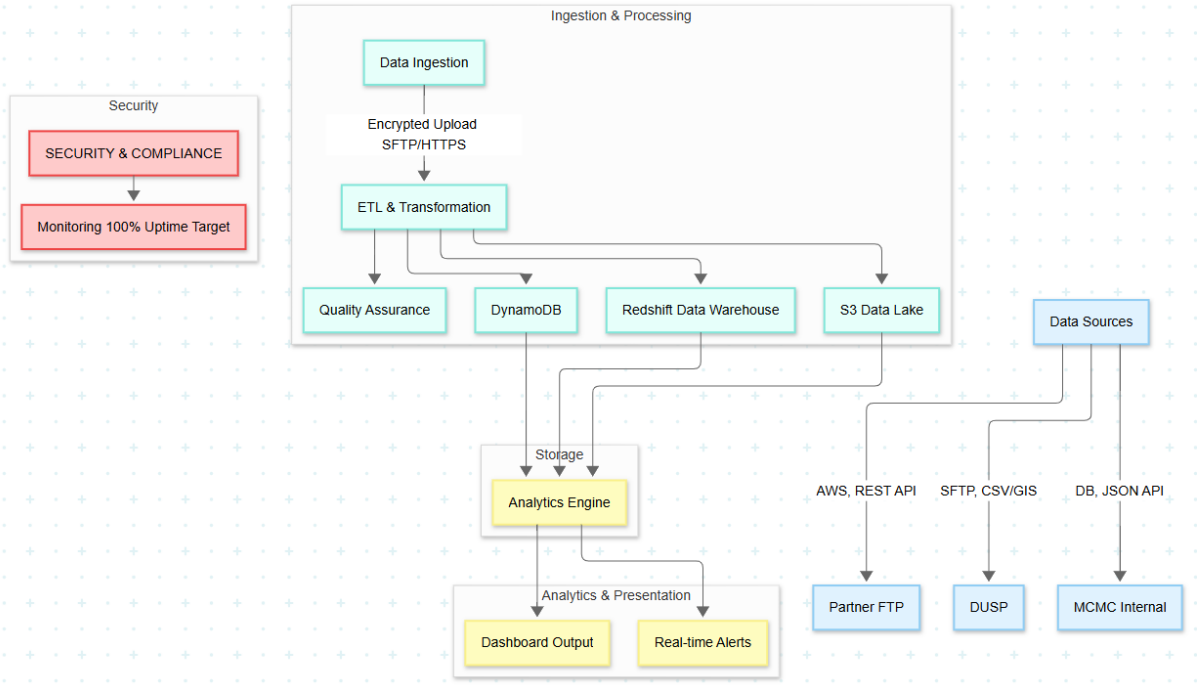


1.2 Core Architecture Components

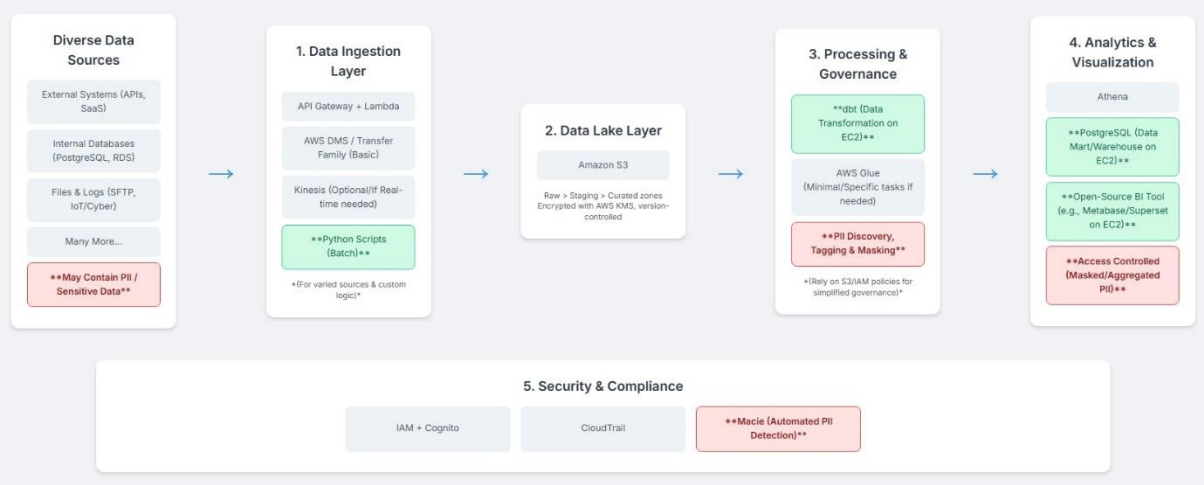
- **Data Ingestion Layer:** Multi-protocol ingestion (SFTP, HTTPS, API, DB), real-time/batch processing, validation, error handling, metadata extraction.
- **Data Processing Layer:** ETL pipelines, data cleansing, normalization, business rule application, quality monitoring.
- **Data Storage Layer:** Raw data lake (S3), structured data warehouse (Redshift), operational data store (DynamoDB), metadata repository.
- **Analytics Layer:** Statistical analysis, machine learning, AI, geospatial analysis, time-series forecasting.
- **Presentation Layer:** Interactive dashboards (Power BI Pro), mobile/web interfaces, API endpoints, automated report generation.

All figures are approximate and derived from the usual data methodology

[Open]



Architecture & Operational Flow



1.3 Core System Components

1. Data Ingestion Layer

- **Multi-Protocol Support**: SFTP, HTTPS, API, Database connections
- **Processing Modes**: Real-time streaming and batch processing
- **Data Validation**: Schema validation, quality assurance, error handling
- **Volume Capacity**: 1-2 million data points daily

All figures are approximate and derived from the usual data methodology

2. Data Processing Engine

- **ETL Pipelines**: Extract, Transform, Load with automated workflows
- **Data Quality**: Cleansing, normalization, business rule application
- **Technologies**: AWS Glue, Step Functions, Apache Spark
- **Monitoring**: Real-time job tracking and alerting

3. Storage Architecture

- **Data Lake**: Raw data storage in Amazon S3
- **Data Warehouse**: Structured analytics in Amazon Redshift
- **Operational Store**: Real-time processing with DynamoDB
- **Security**: AES-256 encryption at rest and TLS 1.3 in transit

4. Analytics & Visualization

- **Primary Platform**: Microsoft Power BI Pro
- **Advanced Analytics**: R/Python for statistical analysis
- **Machine Learning**: AWS SageMaker for predictive models
- **Mobile Access**: Responsive dashboards for executive use

2.0 Detailed Data Flow Architecture

Data Sources Integration:

DUSP (Department of Urban and Spatial Planning) Integration:

- Connection Type: Secure SFTP with certificate-based authentication
- Data Format: CSV, Excel, Geospatial files (SHP, KML)
- Transfer Schedule: Daily batch uploads at 2:00 AM
- Data Volume: 50,000-100,000 records per day
- Validation: Schema validation, data type checking, completeness verification

Internal Systems Integration:

- Connection Type: Database direct connection (encrypted)
- Data Format: Structured database tables, JSON APIs
- Transfer Schedule: Real-time streaming with 15-minute batch processing
- Data Volume: 200,000-500,000 records per day
- Validation: Business rule validation, referential integrity checks

All figures are approximate and derived from the usual data methodology

Technology Partner Integration:

- Connection Type: RESTful APIs with OAuth 2.0 authentication
- Data Format: JSON, XML, structured data feeds
- Transfer Schedule: Hourly updates with real-time alerts
- Data Volume: 100,000-200,000 records per day
- Validation: API response validation, data freshness checks

Data Processing Pipeline:

Stage 1: Raw Data Ingestion

- Automated file detection and processing
- Data format identification and parsing into parquet format.
- Initial data quality assessment
- Metadata extraction and cataloguing

Stage 2: Data Transformation

- Data cleansing and standardization
- Business rule application
- Data enrichment and augmentation
- Master data management and deduplication

Stage 3: Data Loading

- Structured data warehouse population
- Data mart creation for specific use cases
- Index creation and optimization
- Data partitioning and archiving

Stage 4: Quality Assurance

- Data quality monitoring and reporting
- Exception handling and notification
- Data lineage tracking
- Audit trail generation

All figures are approximate and derived from the usual data methodology

3. Technology Stack Specification

3.1 Cloud Infrastructure (Amazon Web Services)

- **Compute**: EC2, Auto Scaling Groups, Elastic Load Balancing, AWS Lambda.
- **Storage**: S3 (data lake), EBS (database), Glacier (archival), EFS (shared files).
- **Database**: Amazon Redshift (data warehousing), RDS (PostgreSQL, MySQL), DynamoDB (NoSQL), ElastiCache (caching).
- **Analytics**: AWS Glue (ETL), Step Functions (workflow), SageMaker (ML), QuickSight (visualization).
- **Security**: IAM, VPC, WAF, GuardDuty.

3.2 Application Stack

- **Data Integration:** Apache Kafka (streaming), Apache Airflow (orchestration), Talend, Custom Python/Java.
- **Analytics Platform:** Microsoft Power BI Pro (primary visualization), R/Python (statistical analysis), Apache Spark (big data), TensorFlow/PyTorch (ML).
- **Monitoring & Operations:** AWS CloudWatch, Elasticsearch/Kibana, Grafana, PagerDuty.

4. Security Architecture & Compliance

4.1 Security Framework

- **Network Security:** VPC, VPN, Network ACLs, AWS Shield (DDoS).
- **Data Security:** Encryption at rest (AES-256), encryption in transit (TLS 1.3), AWS KMS, database/column-level encryption.
- **Identity and Access Management:** MFA, RBAC, least privilege, regular access reviews.
- **Compliance and Governance:** PDPA compliance, data classification, audit logging, regular security assessments.

4.2 Data Privacy Implementation

- **PII Protection:** Data masking/tokenization, secure transmission, access logging, retention/purging policies.
- **Consent Management:** Consent tracking, data subject rights, privacy impact assessments, regular compliance audits.

All figures are approximate and derived from the usual data methodology

5. Operational Procedures & Workflows

5.1 Month 1-3 Operations (MVP Phase)

- **Week 1-2:** Infrastructure setup (AWS account, VPC, EC2, DB).
- **Week 3-4:** Data pipeline development (ETL, source connectivity, initial ingestion, basic dashboard).
- **Week 5-6:** Integration testing (end-to-end, performance, security, UAT).
- **Week 7-8:** Deployment and training (production deployment, user training, go-live support, initial optimization).

5.2 Month 4-7 Operations (Advanced Analytics Phase)

- Advanced analytics development (statistical models, ML pipelines, GIS integration, enhanced dashboards).
- System scaling (performance monitoring, capacity planning, security enhancements, additional data sources).

5.3 Month 8-12 Operations (Predictive Analytics Phase)

- Predictive model development (ML model training, real-time prediction, deployment, monitoring).
- Advanced features implementation (real-time alerts, advanced visualization, mobile app, external system integration).

6. Monitoring & Alerting Framework

6.1 System Health Monitoring

- **Infrastructure:** CPU, memory, disk, network performance, database performance, APM.
- **Application:** ETL job success/failure, data quality metrics, dashboard performance, user activity.
- **Security:** Failed logins, unusual access patterns, data access auditing, incident response.

All figures are approximate and derived from the usual data methodology

6.2 Alerting Configuration

- **Critical Alerts (Immediate):** System downtime, data breach, critical ETL failures, DB connectivity.
- **Warning Alerts (4-hour Response):** Performance degradation, data quality issues, high resource utilization.
- **Informational Alerts (24-hour Response):** Scheduled maintenance, capacity planning, backup completion.

7. Backup & Disaster Recovery

7.1 Backup Strategy

- **Data Backup:** Daily automated DB backups, continuous S3 backup (versioning), weekly full system, monthly Glacier archival.
- **Configuration Backup:** Infrastructure as Code (IaC) templates, application/security configuration, documentation.

7.2 Disaster Recovery Planning

- **Recovery Time Objectives (RTO):** Critical (4 hours), Important (24 hours), Non-critical (72 hours).
- **Recovery Point Objectives (RPO**): Critical (1 hour), Important (4 hours), Non-critical (24 hours).
- **Procedures:** Multi-AZ deployment, cross-region backup, automated failover, regular testing.

8. Performance Optimization

- **Database Optimization:** Query optimization, indexing, partitioning, connection pooling, caching.
- **Application Optimization:** Code optimization, caching, load balancing, scaling, profiling.
- **Infrastructure Optimization:** Resource right-sizing, auto-scaling, network optimization, cost optimization.

All figures are approximate and derived from the usual data methodology

[Open]

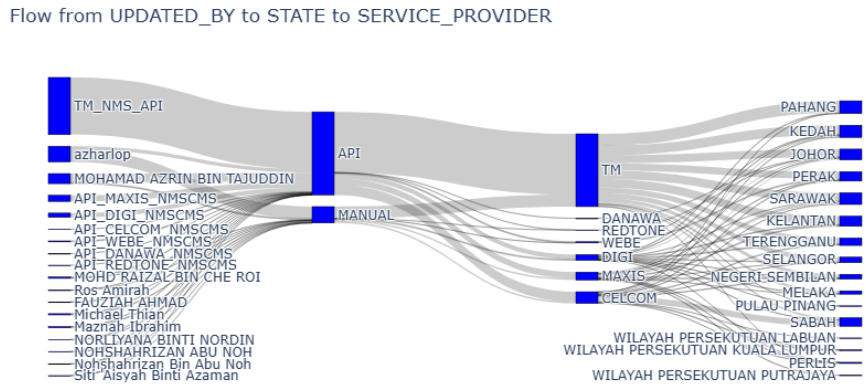
9. Resource Planning & Budget Allocation

9.1 Human Resources (FTE)

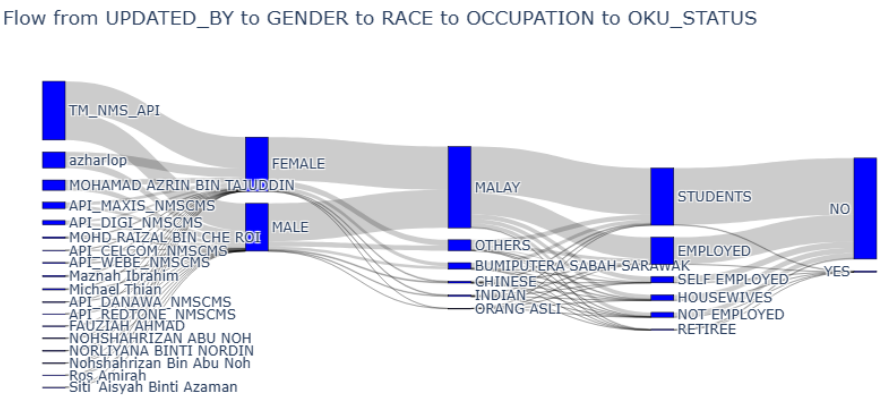
- **Phase 1 (Months 1-3):** Lead Data Scientist (1), Data Engineer (0.5), DevOps Engineer (0.5), Project Manager (0.25).
- **Phase 2 (Months 4-7):** Lead Data Scientist (1), Data Engineers (2), Business Analyst (1), QA Engineer (0.5).
- **Phase 3 (Months 8-12):** Lead Data Scientist (1), Data Engineers (2), ML Engineer (1), UI/UX Designer (0.5).

10. Ambiguous Study from Sample Data

10.1 Data Flow and Tabulation



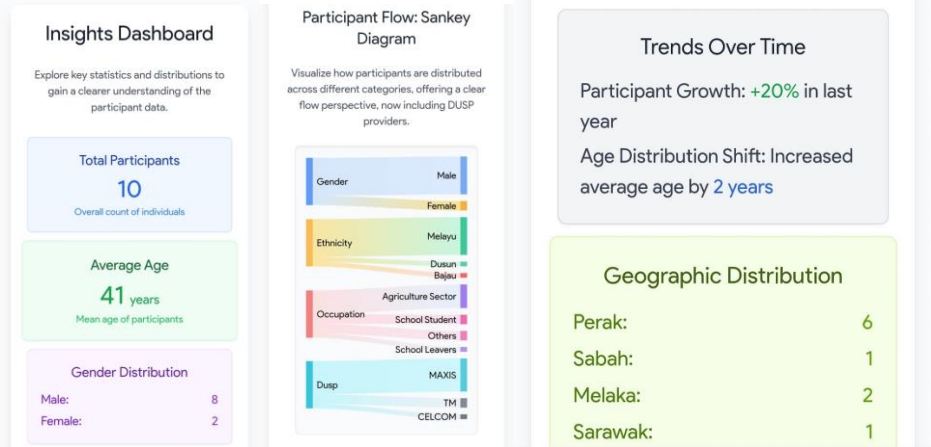
From data source to descriptive statistics



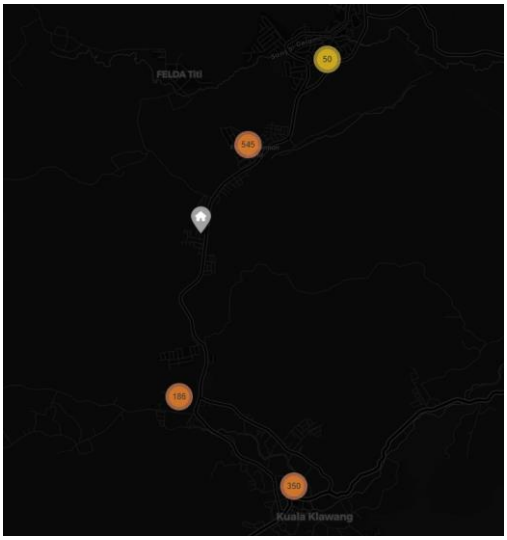
All figures are approximate and derived from the usual data methodology

[Open]

10.2 Sample Data Visualization



Mapping visualization



All figures are approximate and derived from the usual data methodology