SAMARTH

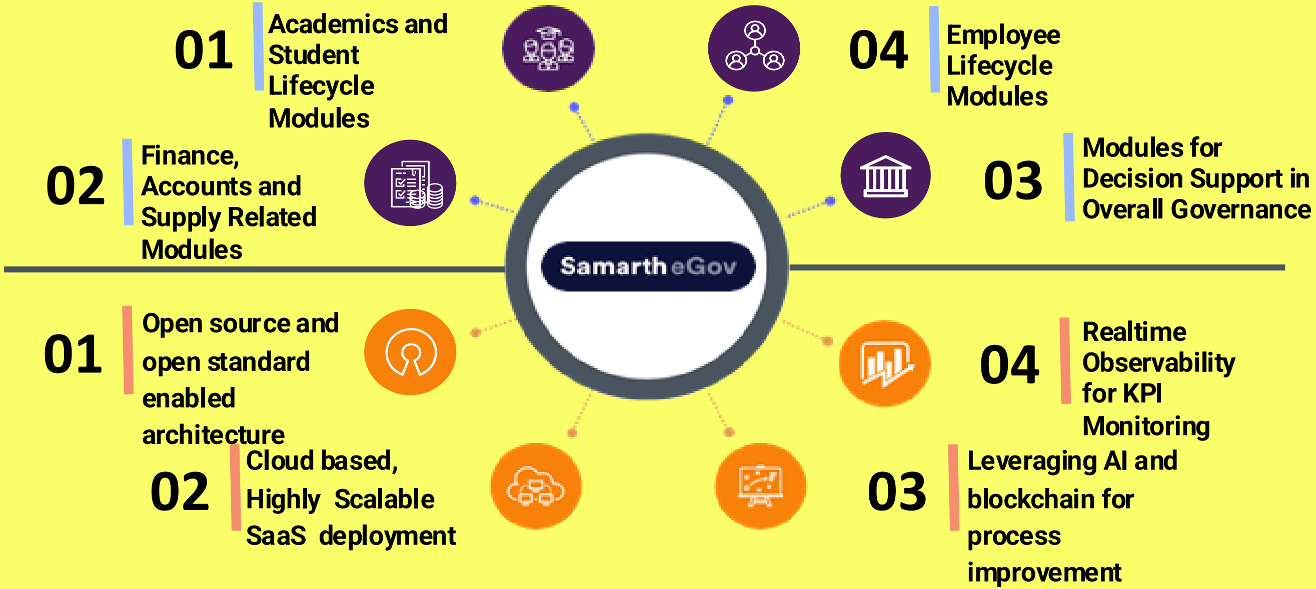
- Designed, Developed and Implemented by IIC, University of Delhi

SAMARTH's USP

Transform your institution with SAMARTH eGov - a powerful, scalable, and future-ready platform built for higher education excellence.

Introduction

Purpose built platform for Higher Education Institutions (HEIs) to deploy e-governance system for administration and management of higher education processes.



Samarth eGov

01 Academics and Student Lifecycle Modules

02 Finance, Accounts and Supply Related Modules

03 Modules for Decision Support in Overall Governance

04 Employee Lifecycle Modules

01 Open source and open standard enabled architecture

02 Cloud based, Highly Scalable SaaS deployment

03 Leveraging AI and blockchain for process improvement

04 Realtime Observability for KPI Monitoring

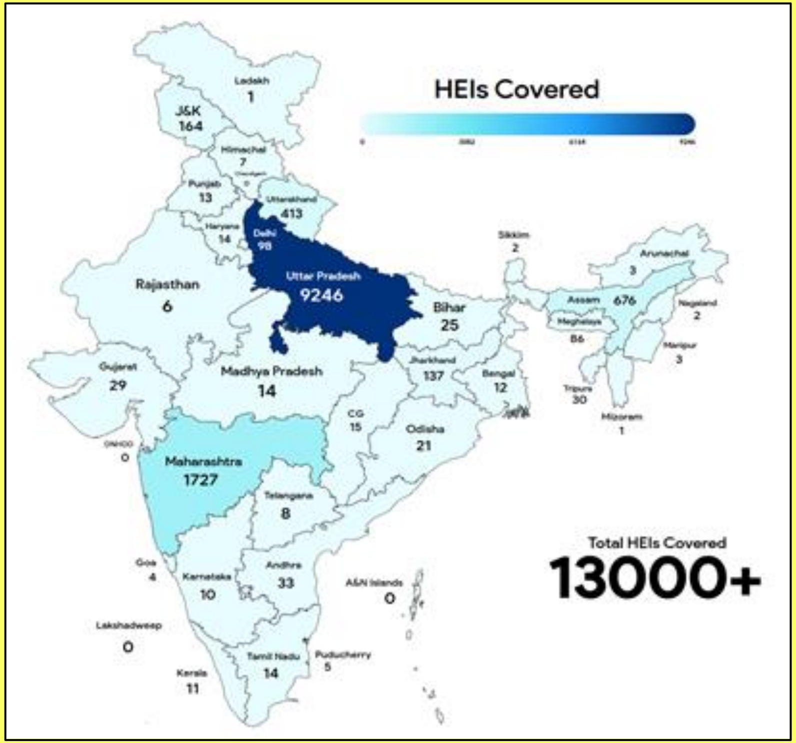
SAMARTH emerges as the point of convergence for higher education policies, compliances and services.

SAMARTH IMPLEMENTATION STATUS – SCALE & SPREAD

Connected **13000+ HEIs** across **32 States/UTs** and **435 Districts/Cities**

STUDENT LIFECYCLE – FROM ADMISSIONS TO CONVOCATION

EMPLOYEE LIFECYCLE – FROM RECRUITMENT TO RETIREMENT



Understanding the SAMARTH Data Landscape

(Confidential)

21.7 M
Records in Student Registry

254 K
Records in Employee Registry

1.24 M
Payroll Statements Processed Online

966 K
Active Students

228 K
Active Employees

1.01 M
Leaves Processed Online

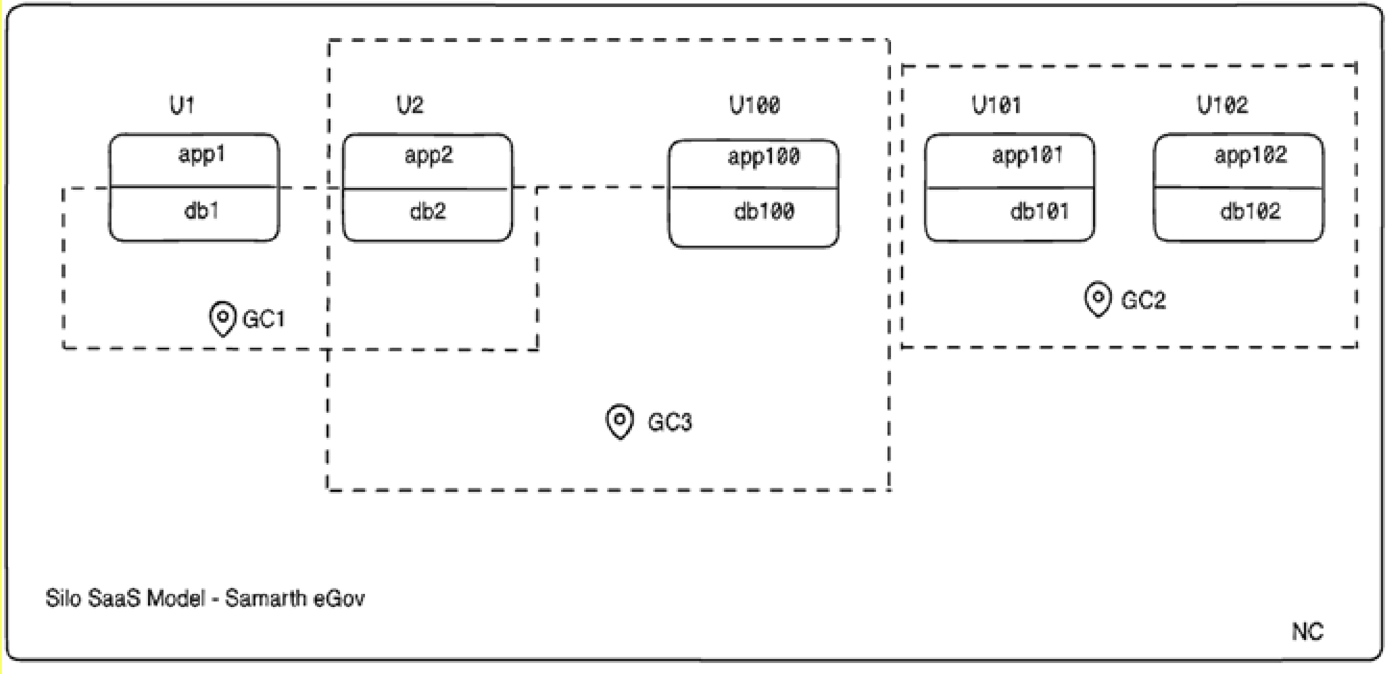
- **Analytical datasets operating at million-level record scale**
- **Continuous, high-volume write amplification workloads**

Dimensions That Matter!

- Nation Top Level [redacted]
- Geographic hierarchies (state, district, region) [redacted]
- Institutional and or [redacted] (university, college, department)
- Temporal dimensio [redacted] (academic year, financial year, seasons, time)
- Staff Attributes (designation, employment nature, service status) [redacted]
- Programme and disciplin [redacted] (stream, course, specializations, elective)
- Demographic attributes (category, reservation, gender)

.. etc

The Dimensional Multiplier: Why Complexity Scales ?



NC - National Cluster
GC* - Geographic Cluster
U* - Individual Subscriber

Symptoms

DB Schemas can vary - Heterogeneous

Computational Latency due to High Dimensionality

Performance Bottlenecks in Deep-Dive Analysis

Excessive CPU and memory consumption due to complex transformations and joins

Heavy on-the-fly aggregations leading to slower response times and unpredictable performance

Journey @ Samarth Analytics Platform

- **Gen1 - Metric-Driven ETL with self defined KPI templates for each subscriber. Additional custom dashboards for NC/GC aggregations**
  - ***Technology -*** *MySQL + Grafana*
  - ***Issues Faced -***
    - Rigid schema, no ad-hoc analysis or correlation
    - Redesign required for every new metric
    - Access provisioning to every end users
    - Cross Subscriber aggregations

Journey @ Samarth Analytics Platform

- **Attempt 1 for Gen2(.v1) – Moving to Data Warehouse**
    - ***Technology –*** *Redshift + Quicksight*
    - ***Issues Faced –***
        - Schema mismatches and modeling errors
        - Data Ingestion overhead
        - High operational effort and cost

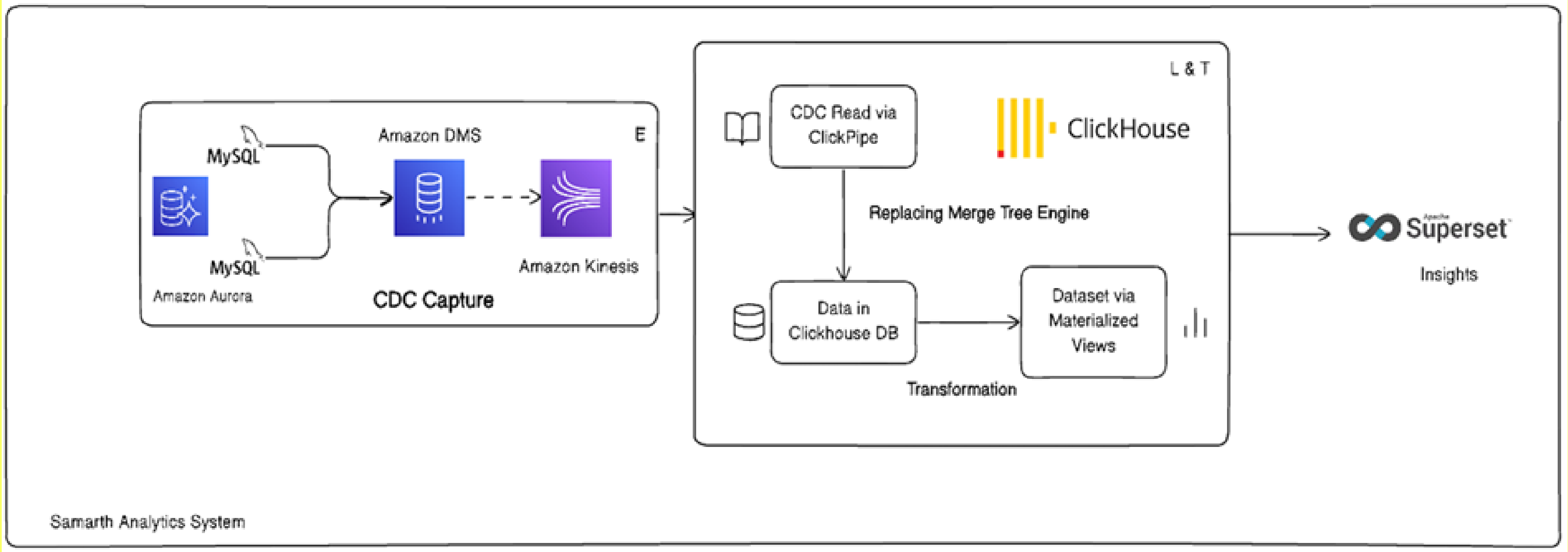
Could not materialize to production

Journey @ Samarth Analytics Platform

- **Attempt 2 for Gen2(.v2) – Data Warehouse using CDC (ELT)**
  - ***Technology –*** *CDC + Clickhouse + Superset*
  - ***Facts –***
    - Schema-Flexible by Design
    - Reduced Pipeline Complexity via Click Pipes
    - Simplifies architecture and lowers operational cost
    - High performance for group aggregations and analytics
    - Fast Execution by using refreshable materialized views

Architecture

End-to-end CDC ingestion via ***Clickpipes*** into ClickHouse, utilizing the **ReplacingMergeTree** engine and **Refreshable Materialized Views** to deliver ***near real-time analytics***



Samarth eGov

ClickHouse Gurgaon Meetup

ClickHouse transformed high-dimensional, high-velocity data into low-latency, cost-efficient analytics at scale

Aspect	v1 – MySQL + Grafana (Classic ETL)	v1.5 – Redshift + QuickSight	v2 – ClickHouse + CDC + Superset
Architecture	Direct queries on source DB with ETL	Centralized warehouse + BI	Streaming CDC-based analytics
Ingestion	Batch ETL	Heavy batch loads	Continuous CDC loading
Schema Handling	Rigid, predefined metrics	Strict schema, load failures on mismatch	Schema-flexible, non-blocking ingestion
Data Model	Siloed, metric-driven	Silo-oriented tables	Unified, analytics-oriented
Failure Impact	Query failures affect dashboards	Load failures block process	Errors isolated in separate tables
Correlation Across Data	Not supported		Fully supported
Query Performance	Slow at scale		Fast group aggregations and analytics
Real-time Capability	No		Near real-time
Visualization Layer	Grafana	QuickSight	Superset
Scalability	Poor		High (designed for scale)

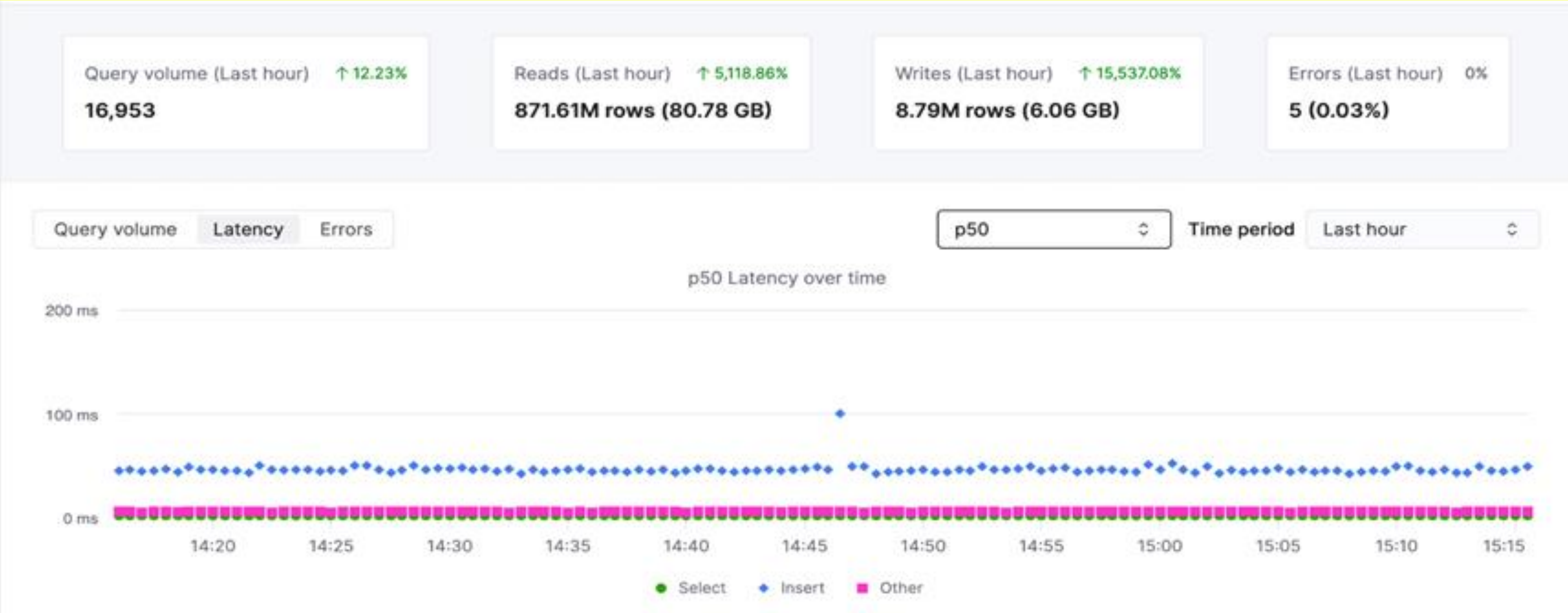
After heavy aggregations and complex fact–dimension joins, e.g. -

Duration Taken: 24.7 seconds

Read Data: 4.85 GB | Written Data: 6.02 GB

Total Rows Materialized: 28.7 million

After heavy aggregations and complex fact–dimension joins, e.g. –



Query volume (Last hour) ↑ 12.23%

16,953

Reads (Last hour) ↑ 5,118.86%

871.61M rows (80.78 GB)

Writes (Last hour) ↑ 15,537.08%

8.79M rows (6.06 GB)

Errors (Last hour) 0%

5 (0.03%)

p50

Time period: Last hour

Thank You!

→ Kunal Sharma

Associate, Data Engineering and AI

SCAN ME