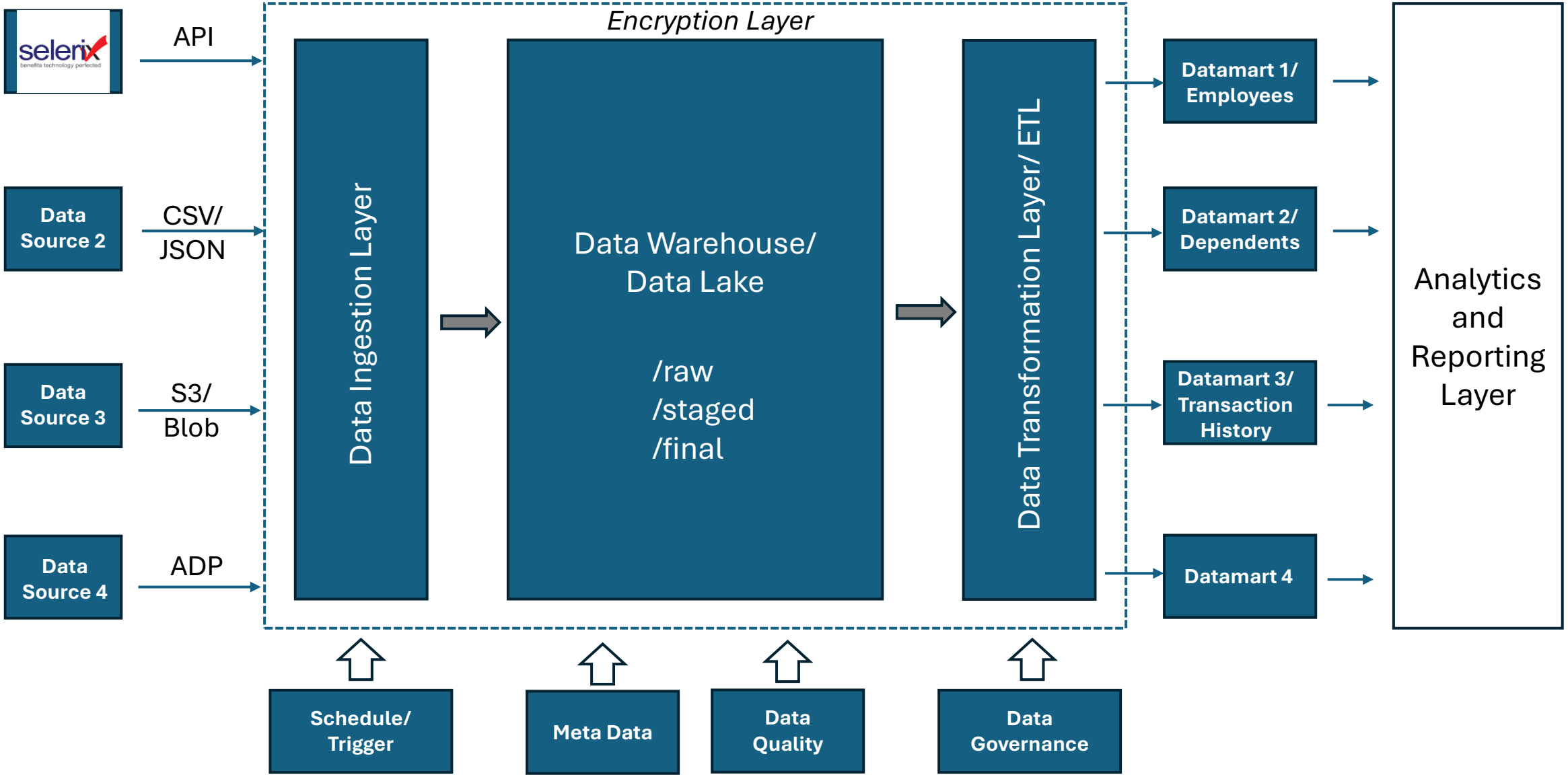
Roadmap to End-to-End Analytics Architecture

By Virinchi Alahari

Contents of the presentation

- High-Level Architecture
- Data Privacy & Security 101
- Change Data Capture of Employee details
- Estimated costs of different system based on the user base
- Low Level Architecture of individual systems
- Implementation Timeline

High Level Architecture



Data Privacy & Security 101

- Some regulators (e.g., FDIC, OCC) may *expect data localization* for **critical infrastructure** or **banking core systems**, even if it's not mandated in the law.
- RBAC (Role Based Access Control)
- Unified Logging System
- Data Encryption
  - At Rest: AES 256 (Highly preferred)
  - In Transit: Use HTTPS only endpoint

Change Data Capture Presentation

Selerix API’s

- **Set Up and Authentication for access to the data**
- **Get Report Task: This API can be used for getting report for Ad-Hoc analysis**
- **Get Template Report Task: This API can be used for incremental refresh of the data and update the dashboards that we already have.**

Reference : https://docs.slrx.io/#405ef211-dbf2-4774-a6dd-ec697a7d0d22

Change Data Capture Presentation

- In the latest report, if an employee decides to add coverage to their kid.
- That change might get lost in the millions of data that we have.
- Created a Prototype to identify mismatches and reconciliation of complex data

```
---------------------------- Log recorded at 2025-06-03 20:04:46.806009----------------------------

--- Differences in sheet: Employees at 2025-06-03 20:04:46.806009---
 Row Column  Original Modified  EID
   2   Last  Campbell   Wilson 6159
   3  First    Brenda     Test 9376
   3   Last Hernandez     Name 9376
```

Data Platform Cost Comparison

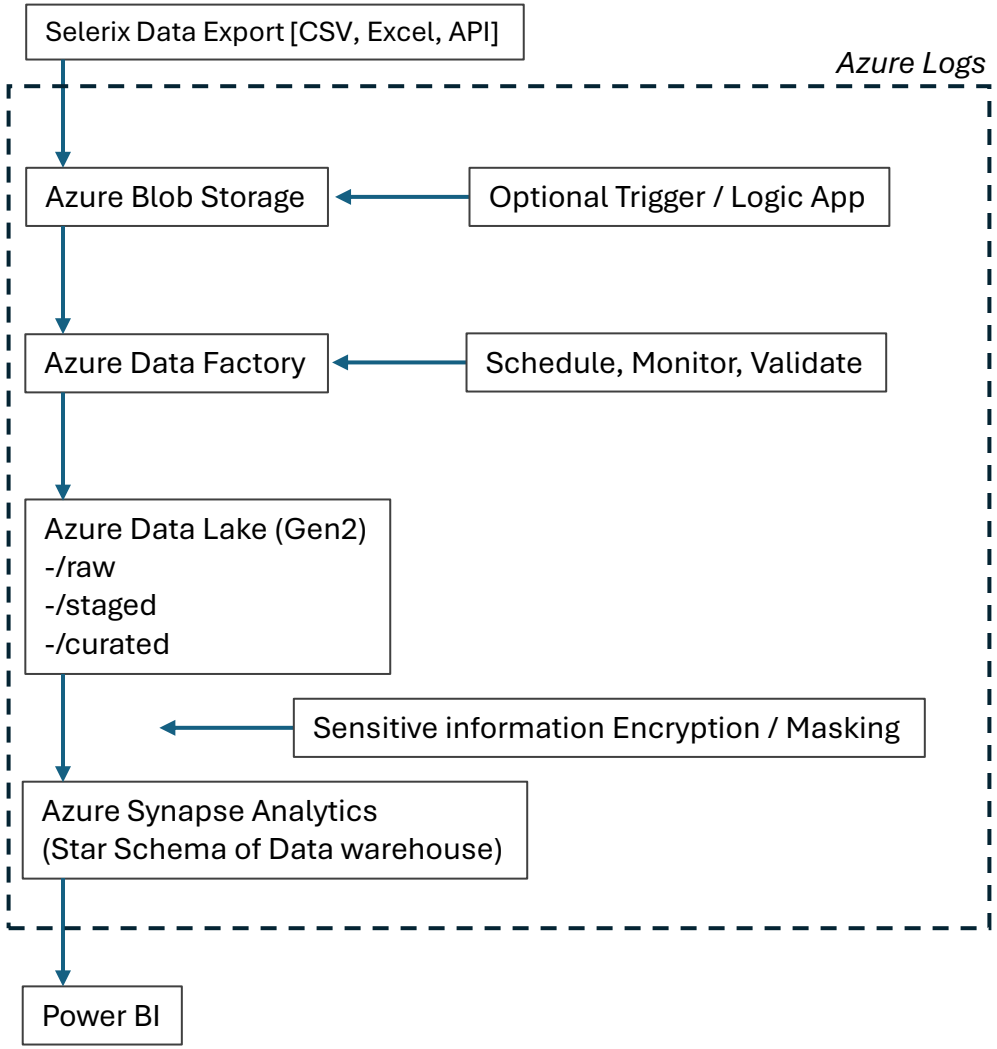
Component	Azure Cloud	Microsoft Fabric	On-Premise
Data Storage	Azure Blob Storage: ~$15–$30/mo	OneLake: Included in F64 SKU	Local storage cost in server price
ETL / Data Pipelines	Azure Data Factory: ~$50–$150/mo	Included (Dataflows, Pipelines)	Python (Free)
Data Warehouse / Engine	Synapse Serverless: ~$100–$250/mo	Lakehouse / Warehouse included	PostgreSQL, Dask, etc. (Free)
Security & Monitoring	Key Vault, Monitor: ~$20–$40/mo	Included (via Purview + Fabric Infra)	Built-in monitoring + manual logging
Hardware / Infra Costs	Cloud-hosted	Cloud-hosted	Physical server: ~$150–$300/month
Total Est. (Hard Cost)	$200–$500/month	$270–$350/month	$200–$500/month (infra only)

Microsoft Fabric (F64–F128): comfortable at ~10–30M/day; needs higher SKU for more
On-Prem: capped by local hardware; 10M/day is doable with a well-tuned system, but scaling is limited without horizontal expansion

Power BI Price Comparison

Tier	Monthly Cost	Storage	Refresh Limits	Sharing
Power BI Free	$0	1 GB	Manual	No sharing
Power BI Pro	$10/user/month	10 GB	8/day	To other Pro users
Power BI PPU	$20/user/month	100 GB	48/day	Premium features

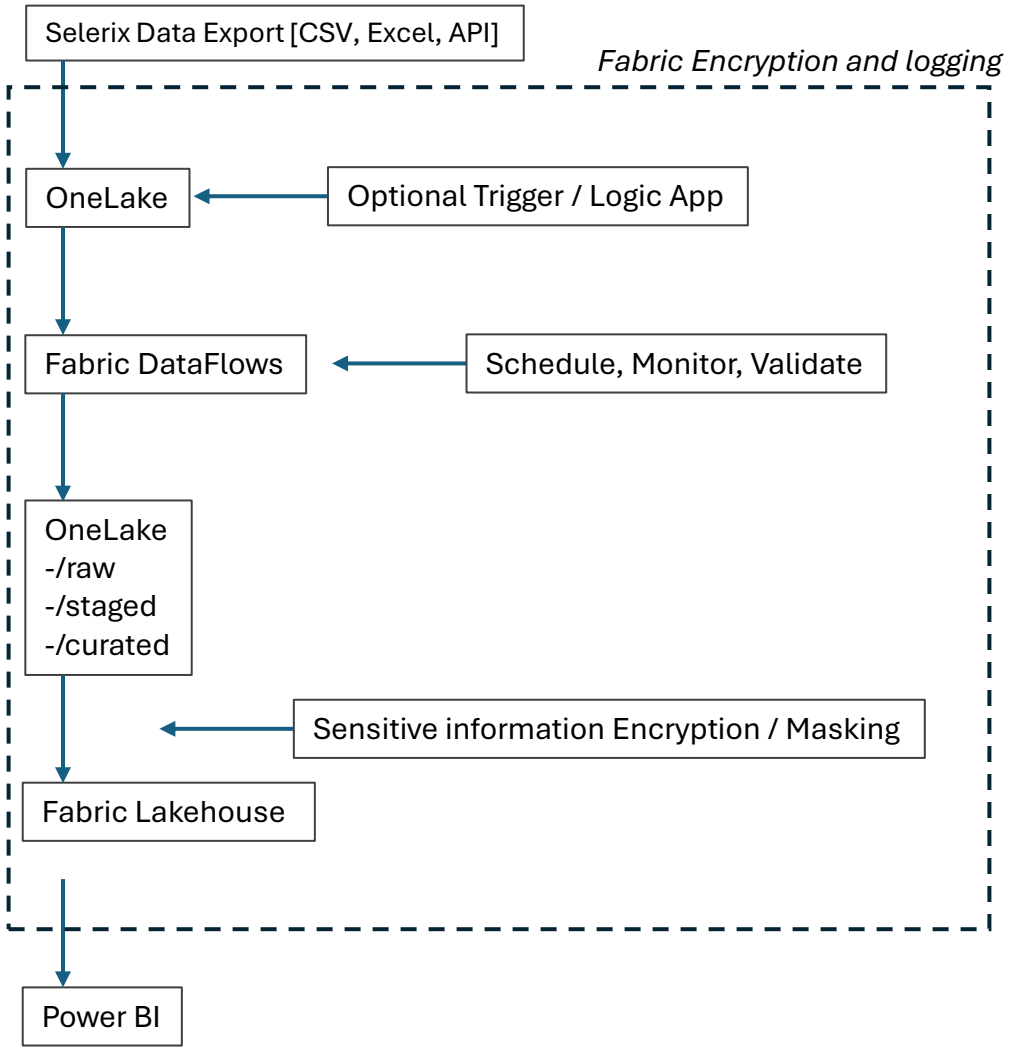
Low-Level Architecture of Modular Azure



Implentation Timeline Approx

Week	Phase	Key Activities
Week 1	Requirements & Environment Setup	- Define KPIs, compliance needs- Set up Azure Blob, ADLS Gen2, Key Vault
Week 2	Data Ingestion Pipeline	- Build ADF pipeline from Selerix exports- Configure triggers or Logic Apps
Week 3	Data Lake Zoning & Transformation	- Create /raw, /staged, /curated layers- Apply data cleaning & masking
Week 4	Synapse Data Warehouse Modeling	- Build star schema (facts/dims)- Load curated data- Apply encryption/masking
Week 5	Power BI Dashboards	- Connect Power BI to Synapse- Build dashboards (demographics, benefits)- Apply RLS
Week 6	Testing & Security Validation	- Full system testing- Validate access, logs, encryption- Finalize SOPs
Week 7	Stakeholder Feedback	- Present solution to business/IT- Collect user feedback on dashboards/process
Week 8	Refinement & Iteration	- Apply improvements- Optimize performance- Finalize documentation

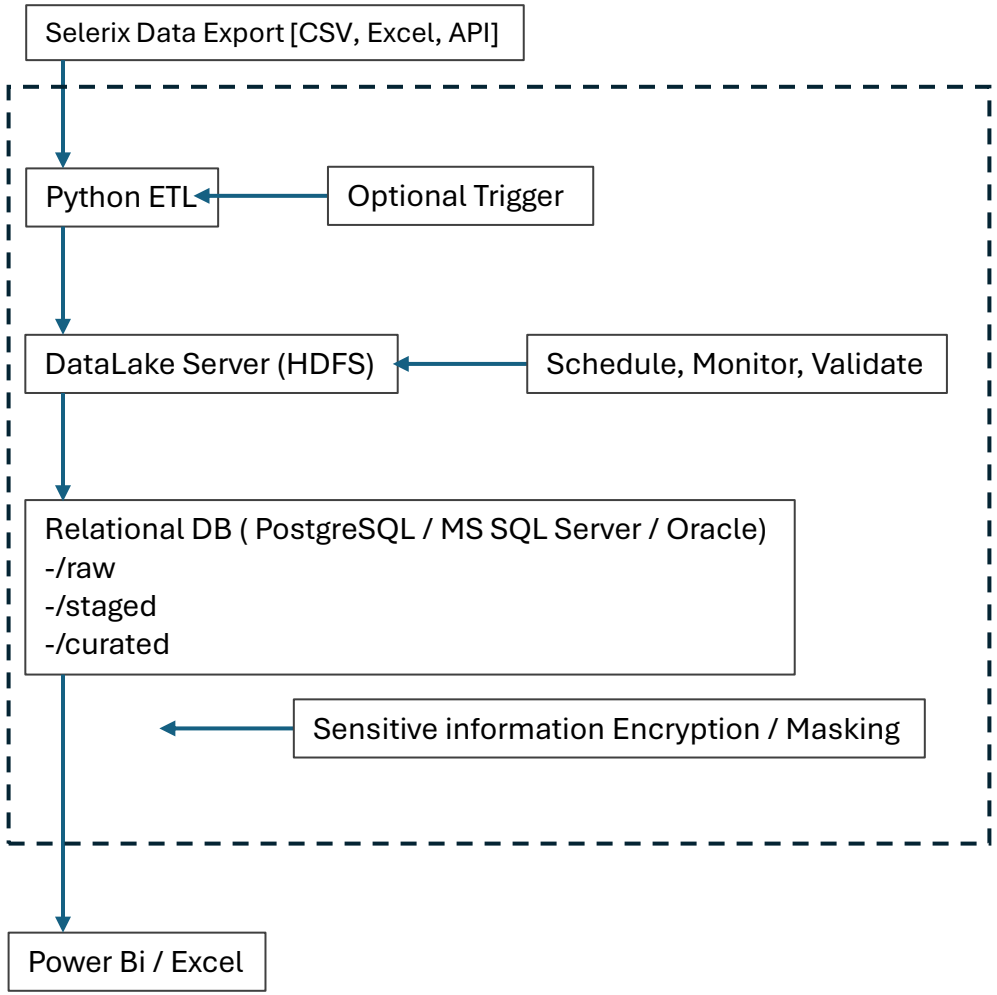
Low-Level Architecture of Azure Fabric



Implentation Timeline Approx

Week	Phase	Key Activities
Week 1	Requirements & Environment Setup	- Define KPIs and reporting goals- Set up Fabric workspace and OneLake containers- Create environments for raw/staged/curated
Week 2	Ingestion with OneLake & Logic Apps	- Connect to Selerix exports (CSV, Excel, API)- Use Logic App or event trigger to upload to OneLake
Week 3	Transformation with Dataflows	- Build **Fabric Dataflows** for cleaning, masking, formatting- Monitor and schedule via Fabric pipeline UI
Week 4	Lakehouse Modeling	- Organize curated tables in **Fabric Lakehouse**- Apply encryption/masking for sensitive fields- Map star schema
Week 5	Power BI Dashboards & Security	- Connect Power BI to Lakehouse- Build dashboards (demographics, benefits)- Apply RLS and dataset refresh logic
Week 6	Feedback & Final Enhancements	- Demo to stakeholders- Apply improvements- Finalize documentation & handoff

Low-Level Architecture of On-Premise Infra



Implentation Timeline Approx

Week	Phase	Key Activities
Week 1	Requirements & Environment Setup	- Finalize KPIs, compliance goals- Set up Linux servers / VMs- Install Python, DBs, and HDFS stack
Week 2	Python ETL Framework	- Build Python scripts for reading CSV/Excel/API- Modularize code (ingest, validate, transform)
Week 3	DataLake Server (HDFS) Integration	- Connect Python to HDFS (via PyArrow, HDFS CLI, or pydoop)- Organize data by zone (/raw, /staged, /curated)
Week 4	Relational DB Design & Load	- Design schemas and tables- Load curated data into PostgreSQL/SQL Server/Oracle- Indexing, partitioning, FK setup
Week 5	Encryption, Masking & Access Control	- Encrypt PII fields- Apply masking (DB-level or in views)- Implement RBAC/user access
Week 6	Visualization Layer	- Connect Power BI or Excel to DB views- Build dashboards (benefits, HR, payroll)- Add filters/security if needed
Week 7	Testing, Tuning & Documentation	- Run end-to-end tests- Validate logs, data flows, query performance- Document setup and SOPs

Thank You