

**Capstone Project:**

# Real Time Financial Investment Insights Platform

## Business Scenario

TVS’s financial division wants to analyze customer investment behavior across multiple avenues (mutual funds, equity, bonds, deposits, gold, etc.). Currently, investment preferences and survey responses are spread across **structured data (investment records), semi-structured data (survey ratings), and unstructured data (free-text objectives/notes)**.

Dummy Data generated or any real data that you can give?

This fragmentation limits the company’s ability to:

* Understand customer investment objectives and risk appetite.
* Identify profitable and at-risk investor segments.
* Provide personalized recommendations.
* Support AI/ML-driven predictive financial advisory.

The capstone project will create a **cloud-native data platform (Azure + Databricks)** to integrate, clean, govern, and analyze investment datasets, providing **real-time dashboards and AI/ML readiness**.

## Key Business Challenges

1. **Data Integration** – Investment preferences, survey responses, and customer demographics are siloed.
2. **Data Governance** – No metadata, lineage, or access control for sensitive financial data.
3. **Data Quality** – Inconsistent survey entries, missing responses, and duplicates.
4. **Customer Insights Gap** – No consolidated view of customer risk appetite vs objectives.
5. **AI/ML Readiness** – No curated datasets for churn prediction, product recommendation, or risk scoring.
6. **Automation & Monitoring** – Manual, Excel-heavy reporting processes with reactive alerts.
7. **Performance & Cost** – Redundant batch pipelines, no real-time processing, inefficient scaling.

This 7th one what is relate to performance and cost cloud cost as we may run redundant batch pipelines and also we may include real time processing and what is inefficient scaling ie due to real time and pipelines scaling issue arises?

## Project Objectives

* **Integrate & Centralize Data** – Build a unified platform for structured (survey responses), semi-structured (investment scores), and unstructured (free-text savings objectives) data.

We will be given datasets great expectation --- is a library helps to do the validations.

Lineage – source and target

* **Enforce Governance & Quality** – Apply data validation, lineage, and secure access controls.
* **Enable Customer Segmentation** – Standardize features to segment investors by age, gender, risk factor, and product preference.
* **Provide Real-Time Dashboards** – Deliver Power BI dashboards showing investor behavior, preferred avenues, and savings objectives.
* **AI/ML Enablement** – Prepare ML-ready datasets for churn prediction, financial product recommendations, and risk profiling.

What are the financial products that you want to recommend based on the analysis of risk profiles?

* **Automate Pipelines & Alerts** – Orchestrate ingestion and processing with ADF + Databricks, with proactive alerts.

How to orchestrate ingestion for this I need to have new data coming from source time to time?

* **Optimize Performance & Cost** – Use Delta Lake partitioning, incremental ingestion, and auto-scaling clusters.

What is incremental ingestion and auto scaling clusters?

## Data Dictionary (Key Fields)

|  |  |
| --- | --- |
| **Field** | **Description** |
| gender | Gender of the respondent |
| age | Age of the respondent |
| Investment\_Avenues | Whether the person invests across different financial avenues (Yes/No) |
| Mutual\_Funds | Preference score for Mutual Funds (scale of 1 to 7) |
| Equity\_Market | Preference score for Equity Market (scale of 1 to 7) |
| Debentures | Preference score for Debentures (scale of 1 to 7) |
| Government\_Bonds | Preference score for Government Bonds (scale of 1 to 7) |
| Fixed\_Deposits | Preference score for Fixed Deposits (scale of 1 to 7) |
| PPF | Preference score for Public Provident Fund (scale of 1 to 7) |
| Gold | Preference score for Gold as investment (scale of 1 to 7) |
| Stock\_Marktet | Whether the respondent invests in stock market (Yes/No) |
| Factor | Key factor influencing investment (e.g., Returns, Risk, Locking Period) |
| Objective | Investment objective (e.g., Capital Appreciation, Income, Growth) |
| Purpose | Overall saving purpose (e.g., Wealth Creation, Savings for Future) |
| Duration | Investment horizon (e.g., 1-3 years, 3-5 years) |
| Invest\_Monitor | Frequency of monitoring investments (e.g., Daily, Weekly, Monthly) |
| Expect | Expected returns in percentage band (e.g., 20%-30%) |
| Avenue | Primary investment avenue used (e.g., Mutual Fund, Equity, Fixed Deposits) |
| What are your savings objectives? | Self-declared savings goal (e.g., Retirement Plan, Education, Health Care) |
| Reason\_Equity | Reason for investing in Equity (e.g., Capital Appreciation, Dividend) |
| Reason\_Mutual | Reason for investing in Mutual Funds (e.g., Better Returns, Fund Diversification) |
| Reason\_Bonds | Reason for investing in Bonds (e.g., Safe Investment, Assured Returns) |
| Reason\_FD | Reason for investing in Fixed Deposits (e.g., Fixed Returns, Risk Free) |
| Source | Source of investment advice or information (e.g., Internet, Financial Consultant) |

## Tasks

### 1. Data Ingestion & Integration

* Collect **customer demographics, investment choices, and ratings** into Azure Blob.
* Stream updates (e.g., new surveys or financial advisor logs) via Event Hub into Databricks.
* Capture **free-text savings objectives** in Cosmos DB.

**Outcome:** Unified financial dataset (structured + semi-structured + unstructured).

So we need to create data if we don’t find any appropriate data source or collection things?

Then only we can use event hub for stream updates?

What is free text saving objectives in cosmo DB?

### 2. Data Transformation & Quality Checks

* Use Databricks Delta Lake for cleaning and standardization.
* Apply quality checks:
  + Normalize survey scales (1–7).

--- means 7 high risk and 1 low risk?

* + Deduplicate entries.tw
  + Validate against expected ranges (age, income brackets, % returns).

---- means categorization according to age, income brackets and returns?

* Feature engineering:
  + Risk Index (based on equity vs FD preference).
  + Investor Confidence Score (based on savings objectives + consistency).

--- means extracting features based on the data paramenters or creating some ananlysis or formula to standardise?

**Outcome:** High-quality curated dataset ready for analysis.

### 3. Governance & Compliance

* Use Unity Catalog for **lineage tracking, role-based access, and metadata management**.
* Apply encryption and access rules for sensitive attributes (income, objectives).
* Maintain audit logs for compliance with financial regulations.

--- whenever we ingest different data ie structured, unstructured and semistructured then we we get all metadata and lineage throught azure inbuilt feature ?? and

--- encrypting personal info like account details and investment detail sand all?

--- running audit time to time if are matching the standards of data privacy acc to financial regulation?

**Outcome:** Trusted, secure financial dataset.

### 4. Curated Data for Analytics & ML

* Store customer-level curated datasets in Synapse SQL for reporting.
* Use Cosmos DB for text-based objectives like *“Retirement Plan”, “Health Care”, “Wealth Creation”*.
* Register features (Risk Index, Confidence Score) in **Databricks Feature Store** for ML use cases.

**Outcome:** Structured + ML-ready data layers.

Make me understand this?

### 5. Real-Time Dashboards

* Create **Power BI dashboards** to visualize:
  + Investment preferences by gender, age, and risk factor.
  + Popular savings objectives (e.g., Retirement vs Wealth Creation).
  + Trends in expected returns (10%–20%, 20%–30%, etc.).
  + Sources of financial influence (Internet, TV, Financial Consultants).

**Outcome:** Stakeholders get real-time insights into investor behavior.

### 6. AI/ML Readiness

* Use curated datasets in Databricks ML to:
  + Predict investor churn (based on dissatisfaction signals).
  + Recommend products (e.g., Equity vs FD based on risk profile).
  + Forecast demand for financial avenues by segment.

**Outcome:** TVS gains predictive analytics capabilities.

--- how to know the first one whether one is staying or not based on what explain?

### 7. Automation & Monitoring

* Orchestrate pipelines in **ADF** (batch) + Databricks Jobs (real-time).
* Configure alerts (e.g., if new survey data is missing or inconsistent).
* Implement **self-healing pipelines** with retry/fallback logic.

**Outcome:** Automated, reliable, and monitored workflows.

### 8. Performance & Cost Optimization

* Partition data by **age group, investment type, and survey batch**.
* Use incremental ingestion (only new or updated survey records).
* Apply auto-scaling for Databricks compute clusters.
* Monitor costs in Azure Cost Management.

**Outcome:** Efficient and cost-optimized pipelines.

## Roadmap

* Ingest financial survey data (CSV/JSON) → ETL basics.
* Apply validation and lineage logging → governance awareness.
* Standardize and enrich features (Risk Index, Confidence Score).
* Create investor dashboards in Power BI.
* Train a simple ML model (e.g., churn prediction) in Databricks.
* Automate with ADF + monitoring and alerts.