**Author**: Munei Tshamano

# Delivery I: Azure/Fabric-Based Data Platform for Social Media Insights

This document outlines a proposed data platform architecture designed to meet Sanlam's need for real-time social media insights. The goal is to enable business units to analyze public sentiment and trending topics around insurance products, supporting data-driven decision-making and digital-first customer engagement.

The solution uses Microsoft Azure and Microsoft Fabric technologies to ensure scalability, reliability, and compliance, while handling both batch and streaming data sources.

**Document Structure**

This document is divided into two parts:

1. **Architecture Overview**
   A breakdown of each pipeline stage, the technology used, and examples.
2. **Architecture Process Flow Diagram**
   A visual representation of the pipeline stages.

## 1. Architecture Overview

The architecture consists of seven key stages:

- Data Ingestion
- Data Processing
- Data Storage
- Data Retrieval & Display
- Orchestration & Scheduling
- Scalability and Reliability
- Data Security & Compliance

### 1. Data Ingestion

**Technology Used**: Azure Event Hubs

Ingests real-time data from social media APIs like Twitter/X using Azure Functions to push data into Event Hubs.

**Example**: Streaming tweets containing keywords such as 'insurance', 'policy', 'claim' etc.

### 2. Data Processing

**Technology Used**: Apache Spark in Microsoft Fabric

Processes data using Spark notebooks for deduplication, null handling, and NLP tasks like sentiment analysis and topic modeling.

**Example**: Classify tweets as positive, negative, or neutral; extract trending topics using LDA.

### 3. Data Storage

**Technology Used**: Microsoft Fabric Lakehouse (Delta Tables)

Stores raw and processed data in Delta Lake format with partitioning by platform, date, and language.

**Example**: Store tweets from Twitter on 2025-06-01 in English under /twitter/2025-06-01 /en/

### 4. Data Retrieval & Display

**Technology Used**: Power BI with Fabric Warehouse

Visualizes insights using dashboards connected to Fabric Warehouse for structured querying.

**Example**: Dashboard showing sentiment trends over time and top influencers by engagement.

### 5. Orchestration & Scheduling

**Technology Used**: Microsoft Fabric Pipelines

Schedules for ingestion and processing jobs with triggers for real-time and batch execution.

**Example**: Daily ingestion of Financial Sanctions List and hourly sentiment scoring of tweets.

### 6.Scalability and Reliability

**Technology Used**: Azure Monitor, Log Analytics, and Auto-scaling in Event Hubs and Spark.

Ensures the platform can handle increasing volumes of data without performance degradation. Azure Event Hubs and Spark pools automatically scale based on load. Geo-redundant storage in Data Lake and Lakehouse ensures availability, while Azure Monitor and Log Analytics provide real-time visibility into system health and performance.

**Example**: During a viral social media event, the system scales up to ingest and process thousands of posts per hour, while monitoring dashboards alert the team to any anomalies or bottlenecks.

### 7. Data Security & Compliance

**Technology Used**: Azure Purview & Azure Active Directory

Ensures PII protection, data lineage, and role-based access control.

**Example**: Mask user handles and locations in tweets before storage; restrict access to sensitive data.

**Author**: Munei Tshamano

## Analytics & Insights Examples

This section outlines the analytics and insights that can be generated from the data platform's outputs:

- **Trending Topics:**

Real-time keyword frequency and clustering using Spark NLP and LDA.

- **Sentiment Analysis:**

Classify tweets as positive, negative, or neutral using pretrained sentiment models.

- **Influencer Detection:**

Identify users with high engagement using follower count and retweet metrics.

- **Geospatial Trends:**

Map sentiment by region using tweet location metadata.

- **Anomaly Detection:**

Detect spikes in mentions or sentiment shifts using time-series analysis.

- **ML Models**

Train models for churn prediction and campaign targeting using historical engagement data.

## 2. Architecture Process Flow Diagram