Business Requirements Specification (BRS)

USAGM Azure Data Platform Project

Document Version: 1.0

Date: August 18, 2025

Prepared By: Icube Consultancy Services Inc.

Reviewed By: John Smith (Director of Digital Analytics, USAGM)

Approved By: Maria Rodriguez (Data Governance Lead, USAGM)

1. Document Information

1.1 Document Control

Field	Value
Document Title	USAGM Azure Data Platform - Business Requirements Specification
Project Name	USAGM Azure Data Platform Implementation
Document Status	Draft
Classification	Internal Use Only
Distribution List	USAGM Leadership Team, Icube Project Team, IT Stakeholders

1.2 Revision History

Version	Date	Author	Description of Changes
1.0	August 18, 2025	Priya Sharma (Icube PM)	Initial draft based on requirements gathering session

2. Executive Summary

2.1 Project Overview

The USAGM Azure Data Platform project will establish a centralized, cloud-based data engineering infrastructure to replace manual Excel-based reporting processes. The platform will automate data ingestion from seven key sources (Emplifi, Voltron, Pangea, Bluesky, Blubrry, Threads, Adobe), implement medallion architecture data lakes, and provide governed analytical datasets for organizational decision-making.

2.2 Business Justification

Current manual processes create significant operational inefficiencies, taking days to generate reports with no audit trails or data lineage. The organization lacks a single source of truth, leading to inconsistent metrics across departments. This data platform will eliminate manual data extraction, establish comprehensive data governance, and enable scalable analytics infrastructure.

2.3 Success Criteria

- Automated daily/weekly/monthly data ingestion from all seven sources
- Establishment of medallion architecture (Bronze, Silver, Gold) data layers
- Implementation of comprehensive data lineage and governance framework

- Reduction of report generation time from days to hours
- Scalable pipeline framework for future data source additions

2.4 Investment Summary

Category	Amount	Justification
Azure Infrastructure	$180,000/year	Data Factory, Databricks, Synapse, Data Lake storage
Development & Implementation	$450,000	Pipeline development, testing, deployment
Training & Knowledge Transfer	$75,000	Team upskilling and documentation
Total Investment	**$705,000**	**18-month ROI through operational efficiency**

3. Current State Analysis

3.1 Current Data Architecture Pain Points

Manual Data Extraction: Teams manually access Emplifi, Voltron, Pangea, Bluesky, Blubrry, Threads, and Adobe platforms individually, downloading files and reports without automation or standardization.

Excel-Based Processing: All data consolidation happens in Excel with manual copy-paste operations, creating version control issues and human error opportunities.

No Data Lineage: Once data enters Excel, there's no traceability of data origins, transformations, or validation processes.

Inconsistent Metrics: Different departments calculate similar metrics differently, leading to conflicting numbers in organizational reports.

Scalability Issues: Each new data source requires manual process development, making growth increasingly difficult to manage.

3.2 Technical Debt Assessment

- No centralized data storage or processing infrastructure
- Absence of automated data quality validation
- Lack of standardized data transformation processes
- No disaster recovery or business continuity for data operations
- Limited ability to handle increasing data volumes

4. Target State Architecture

4.1 Azure Data Platform Design

Data Ingestion Layer: - Azure Data Factory (ADF) for orchestration and scheduling - API connectors for Emplifi, Voltron, Pangea, Threads, Bluesky - File-based ingestion for Adobe exports and Blubrry data - Incremental loading strategies to manage API rate limits

Data Storage Layer: - Azure Data Lake Gen2 with medallion architecture - Bronze Layer: Raw data as extracted from sources - Silver Layer: Cleaned, validated, and standardized data - Gold Layer: Business-ready, curated analytical datasets

Data Processing Layer: - Azure Databricks with PySpark for transformations - Schema validation and drift detection - Deduplication and data quality enforcement - Incremental processing with change data capture

Data Serving Layer: - Azure Synapse Analytics as enterprise data warehouse - Optimized for reporting and analytical workloads - Support for both batch and interactive queries

Governance & Monitoring: - Azure Monitor and Log Analytics for observability - Data lineage tracking through Synapse - Automated alerting for pipeline failures and data quality issues

4.2 Data Flow Architecture

```
Source Systems → ADF → Data Lake (Bronze) → Databricks
→ Data Lake (Silver) → Databricks → Data Lake (Gold)
→ Synapse → Power BI
```

5. Data Source Requirements

5.1 Primary Data Sources

Source	Data Type	Refresh Frequency	Integration Method	Volume	Criticality
Emplifi	Social media analytics	Daily	REST API	10,000 records/day	High
Voltron	Social listening data	Daily	REST API	15,000 records/day	High
Pangea	Content management metrics	Weekly	REST API	5,000 records/week	Medium
Bluesky	Social platform data	Weekly	API/Web scraping	2,000 records/week	Medium
Blubrry	Podcast analytics	Weekly	File export	500 records/week	Low
Threads	Social media data	Weekly	API integration	3,000 records/week	Medium
Adobe	Web analytics	Monthly	File export (CSV)	100,000 records/month	High

5.2 Data Source Integration Specifications

API-Based Sources (Emplifi, Voltron, Pangea, Threads, Bluesky): - OAuth 2.0 authentication where available - Rate limit handling with exponential backoff - Incremental data extraction using timestamps/pagination - JSON response parsing and schema validation - Error handling and retry mechanisms

File-Based Sources (Adobe, Blubrry): - Secure file transfer protocols (SFTP/HTTPS) - Automated file pickup from designated locations - CSV/Excel parsing with column mapping - File archival and retention policies

5.3 Data Quality Requirements

Data Source	Completeness	Accuracy	Timeliness	Consistency
Emplifi	98% of expected records	99% valid metrics	Within 4 hours	Platform standardization
Voltron	95% of expected records	97% valid data	Within 4 hours	Consistent categorization
Adobe	99% of expected records	98% valid sessions	Within 24 hours	Temporal consistency
Others	90% of expected records	95% valid data	Within SLA	Format validation

6. Data Pipeline Requirements

6.1 Ingestion Pipeline Specifications

Daily Pipelines (Emplifi, Voltron): - Execution time: 2:00 AM EST daily - Incremental extraction based on last successful run timestamp - Data validation and schema enforcement - Automatic retry on failure (3 attempts with exponential backoff) - Success/failure notification to operations team

Weekly Pipelines (Pangea, Bluesky, Threads, Blubrry): - Execution time: Sunday 1:00 AM EST - Full or incremental extraction based on data source capabilities - Cross-platform data standardization - Quality validation before promotion to Silver layer

Monthly Pipelines (Adobe): - Execution time: 1st of month, 3:00 AM EST - Large file processing optimization - Historical data validation and gap analysis - Integration with existing monthly reporting cycles

6.2 Transformation Pipeline Requirements

Bronze to Silver Transformations: - Data type standardization and conversion - Null value handling and default assignment - Duplicate record identification and removal - Data format normalization (dates, currencies, text) - Schema evolution handling for API changes

Silver to Gold Transformations: - Business rule application and metric calculations - Data aggregation for reporting dimensions - Slowly Changing Dimension (SCD) implementation - Cross-source data reconciliation - Performance optimization for analytical queries

6.3 Error Handling and Recovery

Pipeline Failure Management: - Automatic pipeline restart for transient failures - Dead letter queue for unprocessable records - Quarantine area for data quality failures - Escalation procedures for repeated failures - Manual intervention processes for critical issues

Data Quality Exception Handling: - Configurable quality thresholds by data source - Automated alerts for quality degradation - Business user notification for critical quality issues - Exception tracking and trend analysis

7. Data Governance Framework

7.1 Data Classification

Classification	Examples	Access Controls	Retention Period
Public	Aggregated social media metrics	All users	3 years
Internal	Detailed engagement data	Department users	5 years
Confidential	Individual user behavior	Authorized analysts only	7 years
Restricted	Personal identifiable information	Data stewards only	Legal requirements

7.2 Data Lineage Requirements

- End-to-end data lineage from source systems to reports
- Transformation logic documentation and versioning
- Impact analysis for schema and business rule changes
- Automated lineage discovery and maintenance
- User-friendly lineage visualization tools

7.3 Data Stewardship Model

Data Owners: Department heads responsible for data definition and business rules **Data Stewards:** Technical leads responsible for data quality and implementation **Data Custodians:** IT team responsible for infrastructure and security **Data Users:** Analysts and business users consuming data for decisions

7.4 Compliance Requirements

- Federal data handling and security requirements
- Data privacy regulations (where applicable)
- Audit trail maintenance for all data modifications
- Change management processes for governance updates
- Regular compliance assessment and reporting

8. Technical Architecture Requirements

8.1 Azure Services Configuration

Azure Data Factory: - Standard pricing tier with auto-scaling - Git integration for pipeline version control - Monitoring and alerting configuration - Cost optimization through off-peak scheduling

Azure Data Lake Gen2: - Standard LRS storage with lifecycle management - Hierarchical namespace for performance optimization - Access control lists (ACLs) for security - 10TB initial capacity with auto-scaling

Azure Databricks: - Standard tier with cluster auto-termination - Unity Catalog for data governance - Job scheduling and monitoring - Cost optimization through spot instances

Azure Synapse Analytics: - Serverless SQL pools for ad-hoc queries - Dedicated SQL pool (DW100c) for production workloads - PolyBase for data lake integration - Backup and disaster recovery configuration

8.2 Performance Requirements

Component	Requirement	Target	Measurement
Data Ingestion	Daily processing completion	Within 2 hours	Pipeline execution time
Query Performance	Dashboard refresh time	< 30 seconds	Power BI monitoring
Data Latency	Data availability after source update	Within SLA by source	End-to-end monitoring
Throughput	Concurrent user support	50 users	Synapse performance metrics

8.3 Security Requirements

Network Security: - Virtual network integration for all Azure services - Private endpoints for sensitive data access - Network security groups with least privilege access - VPN connectivity for on-premises integration

Identity and Access Management: - Azure Active Directory integration - Role-based access control (RBAC) - Multi-factor authentication for all users - Service principal authentication for automation

Data Protection: - Encryption at rest using Azure Key Vault - Encryption in transit using TLS 1.2+ - Column-level security for sensitive data - Dynamic data masking for non-production environments

9. Monitoring and Observability

9.1 Pipeline Monitoring

Azure Monitor Integration: - Custom metrics for pipeline success/failure rates - Log aggregation for error analysis and troubleshooting - Performance metrics for optimization opportunities - Cost monitoring and budget alerts

Operational Dashboards: - Real-time pipeline status monitoring - Data freshness and quality metrics - System performance and resource utilization - Error trending and root cause analysis

9.2 Alerting Framework

Critical Alerts (Immediate Response): - Pipeline failures affecting daily reporting - Data quality issues exceeding thresholds - Security violations or unauthorized access - Service availability issues

Warning Alerts (Business Hours Response): - Performance degradation trends - Approaching storage or compute limits - Data source connectivity issues - Schema drift detection

9.3 Operational Procedures

Daily Operations: - Pipeline execution status verification - Data quality report review - Performance metrics assessment - Issue escalation and resolution

Weekly Operations: - Trend analysis and capacity planning - Security review and access audits - Cost optimization assessment - Stakeholder reporting

10. Testing Strategy

10.1 Data Pipeline Testing

Unit Testing: - Individual transformation logic validation - Data quality rule verification - Error handling scenario testing - Performance benchmarking

Integration Testing: - End-to-end pipeline execution - Cross-source data reconciliation - API integration validation - Disaster recovery procedures

Data Validation Testing: - Source-to-target data accuracy verification - Business rule implementation validation - Historical data consistency checks - Volume and completeness testing

10.2 Performance Testing

Load Testing: - Peak volume processing capability - Concurrent user access simulation - Query performance under load - Resource utilization optimization

Scalability Testing: - Data volume growth simulation - User base expansion testing - Service scaling behavior validation - Cost projection validation

11. Deployment Strategy

11.1 Phased Implementation

Phase 1 (Months 1-3): Foundation - Azure infrastructure setup - Core ADF and Databricks configuration - Emplifi and Voltron integration (highest priority sources) - Basic monitoring and governance framework

Phase 2 (Months 4-6): Expansion - Adobe integration (largest volume source) - Pangea, Threads, Bluesky integration - Advanced data quality and lineage implementation - Power BI dashboard development

Phase 3 (Months 7-8): Completion - Blubrry integration completion - Production optimization and tuning - User training and knowledge transfer - Go-live and hypercare support

11.2 Risk Mitigation

Technical Risks: - API rate limiting: Implement incremental loading and caching - Data quality issues: Establish comprehensive validation frameworks - Performance concerns: Design for scalability from inception - Security vulnerabilities: Implement defense-in-depth architecture

Operational Risks: - User adoption: Comprehensive training and change management - Budget overruns: Continuous cost monitoring and optimization - Timeline delays: Agile development with iterative delivery - Knowledge transfer: Structured documentation and hands-on training

12. Success Metrics and KPIs

12.1 Technical KPIs

Metric	Target	Measurement	Frequency
Pipeline Success Rate	99.5%	ADF monitoring	Daily
Data Quality Score	95%+	Automated validation	Daily
Query Performance	<30 seconds avg	Synapse metrics	Real-time
Data Freshness	Within SLA	End-to-end monitoring	Real-time

12.2 Business KPIs

Metric	Baseline	Target	Measurement
Report Generation Time	2-3 days	4 hours	User feedback
Data Consistency Issues	20+ per month	<2 per month	Quality reports
New Source Onboarding	2 months	2 weeks	Project tracking
Operational Efficiency	Manual processes	80% automation	Process analysis

13. Glossary

13.1 Data Engineering Terms

- **Medallion Architecture**: Data lake design pattern with Bronze (raw), Silver (cleaned), and Gold (curated) layers
- **ETL/ELT**: Extract, Transform, Load / Extract, Load, Transform data processing patterns
- **Data Lineage**: Documentation of data flow from source to consumption
- **Schema Drift**: Changes in data structure that occur over time
- **Incremental Loading**: Processing only new or changed data since last execution

13.2 Azure Services

- **ADF**: Azure Data Factory - cloud-based data integration service
- **ADLS Gen2**: Azure Data Lake Storage Generation 2 - enterprise data lake solution
- **Databricks**: Analytics platform optimized for Azure cloud services
- **Synapse**: Analytics service that brings together data integration, data warehousing, and analytics

13.3 Business Terms

- **Single Source of Truth**: One authoritative data source for business metrics
- **Data Governance**: Management of data availability, usability, integrity, and security
- **SLA**: Service Level Agreement - commitment to specific performance standards
- **API Rate Limiting**: Restrictions on number of API calls within specified time periods

14. Approval and Sign-off

Role	Name	Signature	Date
Business Sponsor	John Smith	[Digital Signature]	[Date]
Data Governance Lead	Maria Rodriguez	[Digital Signature]	[Date]
Project Manager	Priya Sharma	[Digital Signature]	[Date]
Data Architect	Deepak Kumar	[Digital Signature]	[Date]
Senior Data Engineer	Raj Patel	[Digital Signature]	[Date]

Document End

This Business Requirements Specification serves as the foundation for the USAGM Azure Data Platform implementation. All subsequent technical designs, development activities, and testing procedures must align with the requirements outlined in this document.