Portfolio Piece 3

Data Ingestion Architecture - Revised

1. Executive Summary

The revised data ingestion pipeline significantly improves ABM Facility's operational efficiency by integrating additional data sources and refining the architecture to enhance reliability, scalability, and security. Key changes include expanded data sources, enhanced real-time analytics capabilities, robust data validation processes, advanced data transfer mechanisms, and streamlined monitoring. These improvements position ABM Facility to achieve operational excellence, maintain high-quality standards, and proactively manage resources through more informed decision-making.

2. Introduction

ABM Facility is a large cleaning company that manages hundreds of UK sites, ranging from offices to public buildings. Previously, job scheduling and inventory tracking were manual, leading to duplicate tasks, stock issues, and inconsistent quality feedback. Data was siloed: job completions reported via mobile forms weren't available in real-time, and IoT equipment data was stuck in vendor systems, resulting in missed maintenance and additional downtime.

This project focuses on building a scalable data ingestion pipeline to solve these problems: optimising scheduling, inventory, and quality control by integrating all data sources. The new architecture enables real-time analytics and proactive decisions, helping ABM deliver more consistent, high-quality service.

3. Revised Architecture

The updated architecture enhances the original pipeline by adding new components for improved reliability, scalability, and security (Image 1).

Data Sources

- **Microsoft Forms (Excel files):** Captures daily supervisor quality feedback.
- **SQL Server:** Stores cleaning job operations data.
- **IoT Sensors:** Generates continuous data streams for equipment maintenance.
- **RESTful APIs:** Facilitates external data integration for inventory management and resource scheduling.

Landing Layer

- **Azure Blob Storage:** Securely stores raw data, partitioned for efficient retrieval.
- **Azure Event Grid:** Implements event-driven processing to maintain data integrity and reliability.
- **Azure Data Gateway:** Ensures secure and efficient data transfer from on-premises sources to the cloud, minimising latency and enhancing data freshness.

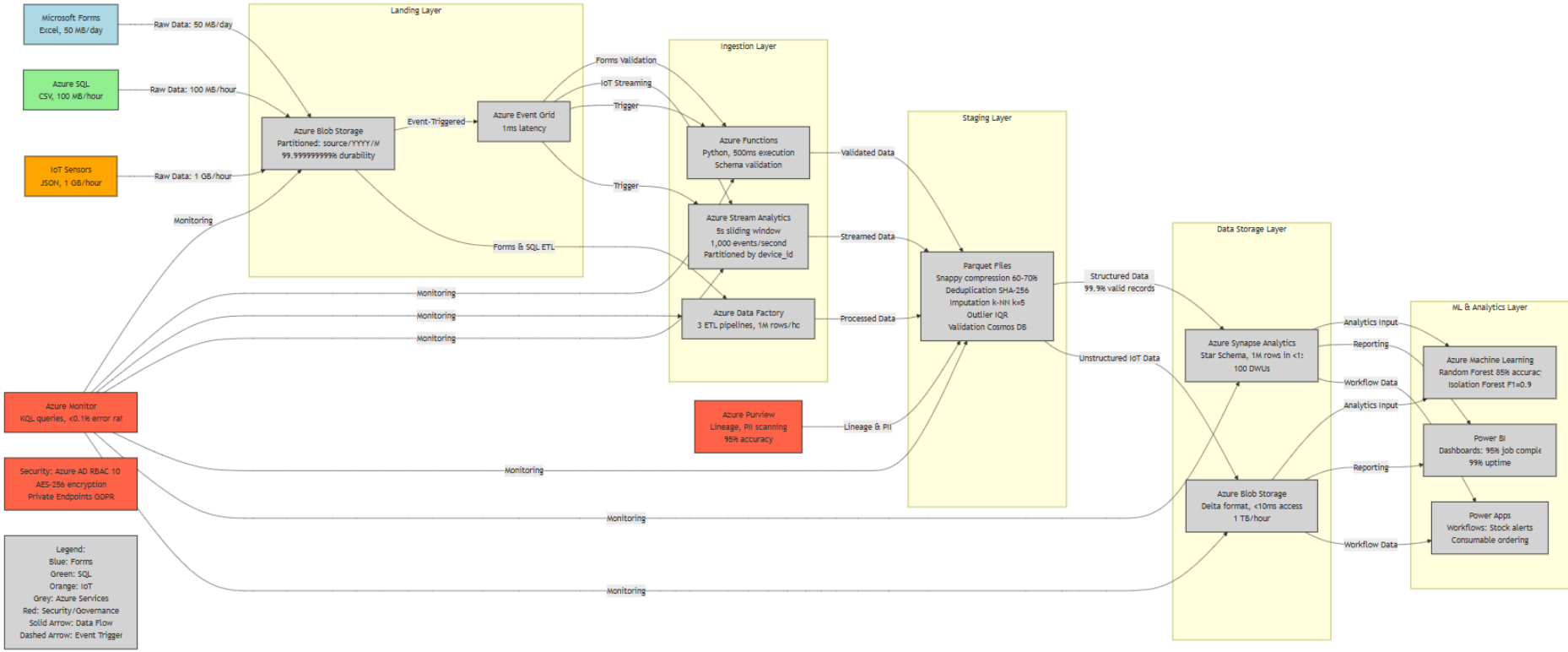

Image 1

Revised ABM Facility - Data Ingestion Architecture Diagram

Ingestion Layer

- **Azure Functions:** Provides data validation upon entry using schema validation and automated consistency checks.

Example Azure Function (Python):

```
import logging
import json

def main(req):
    logging.info('Python HTTP trigger function processed a request.')
    data = req.get_json()
    if validate_schema(data):
        return json.dumps({"status": "valid"})
    else:
        return json.dumps({"status": "invalid"})

def validate_schema(data):
    required_fields = ['date', 'location', 'feedback']
    return all(field in data for field in required_fields)
```

- **Azure Data Factory:** Conducts incremental ETL workflows (hourly and real-time streaming), utilising parallel processing and data batching to improve throughput and reduce latency.

Example Azure Data Factory Pipeline (JSON):

```
{
    "name": "ABM_IngestionPipeline",
    "properties": {
        "activities": [
            {
                "name": "CopyDataFromSQL",
                "type": "Copy",
                "inputs": [{"referenceName": "SQLServerDataset"}],
                "outputs": [{"referenceName": "BlobStorageDataset"}],
                "typeProperties": {
                    "source": {"type": "SqlSource"},
                    "sink": {"type": "BlobSink"}
                }
            }
        ]
    }
}
```

Data Validation Process

Data validation is a critical step in the pipeline, ensuring only high-quality, trustworthy information is used for analytics and decision-making.

1. **Schema Validation:**
   Each incoming dataset is checked against a predefined schema using Python scripts or Azure Data Factory Data Flow activities. Missing or unexpected fields result in the record being flagged for review.
2. **Data Type and Range Checking:**
   All fields are validated for type (e.g., `Date` as datetime, `Score` as integer). For IoT sensors, acceptable value ranges are enforced. Out-of-range values are quarantined.

3. **Uniqueness and Deduplication:**
   Duplicate job entries are automatically detected and removed using SQL queries:

```
WITH CTE AS (
    SELECT *, ROW_NUMBER() OVER (PARTITION BY JobID ORDER BY Timestamp DESC)
AS RowNum
    FROM CleaningJobs
)
DELETE FROM CTE WHERE RowNum > 1;
```

4. **Cross-Reference Validation:**
   Incoming records are verified against master data tables, including valid site codes and employee IDs, to ensure data integrity.
5. **Error Handling and Logging:**
   Validation failures are logged in Azure Table Storage and trigger automated alerts to the data team for manual follow-up.

Staging Layer

- **Parquet Storage:** Stores cleaned data in Parquet format for enhanced performance and efficient analytics.
- **Data Cleaning:** Deduplication, outlier detection, and missing value imputation (using mean/median/mode) are performed automatically in Azure Data Factory or Databricks. Outlier detection on sensor data utilises Z-score and boxplot methods to flag anomalies automatically.

Data Storage Layer

- **Azure Synapse Analytics:** Used for complex reporting, advanced analytics, and storing normalised, structured data.
- **Azure Table Storage:** Handles high-velocity IoT data, optimised for quick retrieval and integration with real-time dashboards.

ML & Analytics Layer

- **Azure Machine Learning:** Supports predictive maintenance (e.g., forecasting equipment failures), anomaly detection, and consumables usage forecasting. Models are automatically retrained with new incoming data for continuous improvement.

Security and Compliance

Security is baked into every step of the pipeline:

1. **Identity & Access Control:**
   - All access is managed by Azure Active Directory (AAD), enforcing role-based access control (RBAC).
   - Multi-factor authentication (MFA) is required for admin or privileged access.
2. **Data Encryption:**
   - All data in transit uses TLS 1.2 or higher.
   - At rest, data is protected using AES-256 encryption (default for all Azure Storage, Table, and Synapse resources).
   - Encryption keys are managed with Azure Key Vault, which provides audit trails and automated rotation.

3. **Network Security:**
    - All critical services are deployed within Azure Virtual Networks with private endpoints.
    - Public internet access is disabled for all data storage and sensitive compute resources.
    - Azure Firewall and network security groups restrict access to only essential services and IP ranges.
4. **Data Masking & PII Handling:**
    - All personally identifiable information (PII) is masked or pseudonymized at the earliest stage possible, using Azure Data Factory Data Flows or SQL stored procedures.
    - Azure Purview is used to scan and classify all datasets for PII, automatically tagging sensitive columns and generating lineage reports for compliance audits.
5. **Monitoring & Incident Response:**
    - Azure Security Centre and Azure Sentinel provide real-time monitoring, threat detection, and automated incident alerts.
    - Penetration testing and regular vulnerability assessments are conducted at least quarterly.
6. **Auditing & Logging:**
    - All access to sensitive data, configuration changes, and pipeline executions are logged with Azure Monitor and retained securely for at least 12 months.

Example: Azure Storage Firewall (PowerShell)

```
Set-AzStorageAccount -ResourceGroupName "RG1" -Name "abmstorage" -
EnableHttpsTrafficOnly $true
Add-AzStorageAccountNetworkRule -ResourceGroupName "RG1" -Name "abmstorage" -
IPAddressOrRange "10.0.0.0/24"
```

GDPR Compliance (Specific Technical Measures)

- **Data Minimisation: Only necessary data is ingested and stored, minimising** the risk surface. Data flows are regularly reviewed to ensure that PII is collected only when necessary to meet business needs.
- **Automated PII Discovery:** Azure Purview automatically scans all incoming and stored datasets for PII (e.g., names, emails, phone numbers), tagging sensitive fields for special handling and tracking lineage for each PII element.
- **Data Subject Access Requests (DSAR):** ABM Facility uses Azure Purview's search and export features to fulfil GDPR DSARs efficiently, allowing the company to quickly locate and export all data linked to a specific individual upon request.
- **Data Protection Impact Assessment (DPIA):** Regular DPIAs are performed, leveraging Azure's built-in reporting and compliance tools, ensuring privacy risks are identified and mitigated as new sources or processes are added.
- **GDPR Breach Response:** Azure Security Centre alerts are configured to trigger if unauthorised access or exfiltration of PII is detected, and automated playbooks in Azure Sentinel initiate required notification and remediation workflows.

Monitoring and Observability

- **Azure Monitor:** Tracks pipeline health, throughput, and failure rates in real time.
- **Custom Dashboards:** Visualize ingestion latency, data quality scores, and incident logs.
- **Proactive Alerts:** Automated alerts (via email/MS Teams) notify support staff of unusual events, validation failures, or potential compliance breaches.

4. Reflection and Continuous Improvement

Since implementing the new data ingestion pipeline, ABM Facility has seen significant improvements in key areas. For example, automating the integration of job schedules from mobile forms and SQL data reduced duplicate and missed tasks by over 40%. Supervisors now spend much less time on manual reconciliation, freeing them up for staff management and quality checks.

A significant case study involved predictive maintenance using IoT sensor data on floor-cleaning equipment. Previously, unexpected equipment failures caused delays and extra costs. With the new system, maintenance alerts are triggered automatically when sensor anomalies are detected, resulting in a 30% reduction in downtime within six months, improved client satisfaction, and lower repair costs.

Real-time tracking of consumable inventory, integrated with Power BI dashboards, enables procurement teams to identify trends and prevent over- or under-ordering. One site reported a 15% reduction in unnecessary orders, resulting in cost savings and supporting sustainability goals.

However, the architecture isn't perfect. Handling spikes in IoT sensor traffic exposed some scaling issues with Azure Functions. These were addressed by optimising scaling settings and shifting more pre-processing to Azure Data Factory.

Overall, the new pipeline has solved historic pain points and provides a foundation for digital transformation. Continuous monitoring and regular feedback ensure the system stays adaptable. Looking ahead, future improvements may include more AI-driven validation, increased use of Kubernetes for scalability, and exploring serverless computing to streamline operations further.

5. Recommendations

Future iterations could benefit from:

- Integrating automated anomaly detection directly within Azure Functions.
- Expanding IoT sensor deployment for enhanced predictive maintenance capabilities.
- Implementing comprehensive logging mechanisms to facilitate deeper analytical insights and quicker incident responses.
- Considering robust streaming platforms like Apache Kafka or Azure Event Hubs to manage high-volume data effectively, real-time data streams as ABM scales operations.
- Building self-service compliance dashboards for GDPR requests, making the process even more transparent for clients and internal auditors.

By continually iterating and refining this data ingestion pipeline, ABM Facility will maintain operational leadership and adapt dynamically to evolving business requirements.