Stakeholder Interview Responses

Email From: Thozamile Mbalo:

Good day, Team,

I trust this email finds you well.

I am busy with the Requirement Analysis of the project. Can you please get the relevant stakeholders to answer the below questions as it will really help us with a better understanding and a better tailored project outcome.

Business Context and Ownership:

1. Who owns the data?
2. What Business Process does it support?
3. Are there any system and data documentation I can leverage?
4. Does a data model and data catalogue exist?

Architecture and Technology Stack:

1. How is the source data stored? (SQL Server, Oracle, AWS, Azure, MongoDB, ...)
2. What are the integration capabilities? (API, Kafka, File Extract, Direct DB, ...)

Extract and Load:

1. Incremental vs. Full Loads?
2. Data Scope and Historical Needs?
3. What is the expected size of the extracts?
4. Are there any data volume limitations?
5. How to avoid impacting the source system's performance?
6. Authentication and authorisation (tokens, SSH keys, VPN, IP whitelisting, ...)

Kind Regards,

Thozamile Mbalo

Data Solutions Architect

Below are the simulated responses from the “stakeholders” for each source system, addressing the questions to inform the requirements analysis.

1. Transactional Data (SQL Server – Orders)

Stakeholder: Sales Operations Team Lead

Business Context and Ownership:

1. **Who owns the data?**
   - The Sales Operations team owns the orders data, managed by the Sales IT group.
2. **What Business Process does it support?**
   - Supports order processing, customer relationship management, and sales performance tracking. Used for revenue reporting, customer segmentation, and product performance analysis.
3. **Are there any system and data documentation I can leverage?**
   - Yes, a database schema diagram and table metadata are available in the internal wiki. Documentation includes column descriptions (e.g., email_pii, product_category).
4. **Does a data model and data catalogue exist?**
   - A logical data model exists, detailing tables like Orders. A partial data catalogue lists key columns but lacks business context for some fields (e.g., total_amount).

Architecture and Technology Stack:

1. **How is the source data stored?**
   - Stored in an on-premises SQL Server 2019 database (SalesDB), hosted on Windows Server.
2. **What are the integration capabilities?**
   - Supports DirectDB connections via ODBC/JDBC, file extracts (CSV), and REST API for specific endpoints (e.g., recent orders). No Kafka support.

Extract and Load:

1. **Incremental vs. Full Loads?**
   - Prefer incremental loads based on order_date to capture new or updated orders daily.
2. **Data Scope and Historical Needs?**
   - Latest 30 days of data is sufficient for analytics; historization not required.
3. **What is the expected size of the extracts?**
   - Daily incremental extract: ~10,000 rows (~5 MB CSV). Full load: ~1M rows (~500 MB).
4. **Are there any data volume limitations?**
   - No strict limitations, but extracts should be scheduled outside business hours (e.g., midnight SAST).
5. **How to avoid impacting the source system's performance?**
   - Use read-only replicas for queries, schedule extracts during low-usage periods, and limit query complexity.
6. **Authentication and authorization?**
   - SQL Server authentication with username/password. IP whitelisting required for external access. No tokens or SSH keys.

2. Clickstream Data (MongoDB)

Stakeholder: Web Analytics Team Lead

Business Context and Ownership:

1. **Who owns the data?**
   - The Web Analytics team owns the clickstream data, managed by the Digital IT group.
2. **What Business Process does it support?**
   - Supports website performance tracking, user behavior analysis, and marketing campaigns. Used to analyze session duration, page views, and product engagement.
3. **Are there any system and data documentation I can leverage?**

- Limited documentation: a MongoDB collection schema in a shared Confluence page, listing fields like session_duration_seconds and product_id.

4. **Does a data model and data catalogue exist?**
    - No formal data model. A basic data catalogue exists for key fields but is incomplete for nested attributes.

Architecture and Technology Stack:

1. **How is the source data stored?**
    - Stored in MongoDB Atlas (cloud-hosted MongoDB) on AWS, in a collection clickstream_sessions.
2. **What are their integration capabilities?**
    - Supports MongoDB API for direct queries, file exports (JSON), and Kafka for real-time streaming. File exports are preferred for batch processing.

Extract and Load:

1. **Incremental vs. Full Loads?**
    - Incremental loads based on timestamp for new sessions daily.
2. **Data Scope and Historical Needs?**
    - Latest 30 days of data for analytics; no historical data required.
3. **What is the expected size of the extracts?**
    - Daily incremental extract: ~20,000 documents (~10 MB JSON). Full load: ~2M documents (~1 GB).
4. **Are there any data volume limitations?**
    - No strict limitations, but large exports may incur costs on MongoDB Atlas.
5. **How to avoid impacting the source system's performance?**
    - Use secondary read replicas, limit query scope (e.g., filter by timestamp), and schedule exports during off-peak hours.
6. **Authentication and authorization?**

- MongoDB Atlas API keys for authentication. IP whitelisting required. No VPN or SSH keys.

3. Inventory Data (CSV)

Stakeholder: Inventory Management Team Lead

Business Context and Ownership:

1. **Who owns the data?**
   - The Inventory Management team owns the data, managed by the Supply Chain IT group.
2. **What Business Process does it support?**
   - Supports inventory tracking, stock replenishment, and warehouse operations. Used for stock level monitoring and supply chain optimization.
3. **Are there any system and data documentation I can leverage?**
   - Basic documentation: a CSV file schema in a shared OneDrive folder, listing columns like warehouse_name and stock_quantity.
4. **Does a data model and data catalogue exist?**
   - No formal data model or catalogue. The CSV schema serves as the primary reference.

Architecture and Technology Stack:

1. **How is the source data stored?**
   - Stored as CSV files in a shared SFTP server, updated daily by an ETL process from the ERP system.
2. **What are the integration capabilities?**
   - File extracts via SFTP. No API or direct database access.

Extract and Load:

1. **Incremental vs. Full Loads?**
   - Full loads daily, as the CSV represents the latest snapshot.
2. **Data Scope and Historical Needs?**

- Latest snapshot only; no historical data required.

3. **What is the expected size of the extracts?**
    - Daily full load: ~5,000 rows (~2 MB CSV).
4. **Are there any data volume limitations?**
    - No limitations, but SFTP bandwidth may be constrained during peak usage.
5. **How to avoid impacting the source system's performance?**
    - Schedule downloads during off-peak hours (e.g., 2 AM SAST) to avoid SFTP server congestion.
6. **Authentication and authorization?**
    - SFTP credentials (username/password). SSH keys preferred for automation. IP whitelisting optional.