# Key Questions to Ask When Connecting Data to a New Warehouse

## 1. Business Context & Data Ownership

Understanding the business side of the data ensures proper alignment with organizational needs:

* • Who owns the data?
* Identify the stakeholders or teams responsible for managing and maintaining the data.
* • What business processes does the data support?
* Understand how the data is used to support operations, decision-making, or analytics.
* • Is there proper system and data documentation?
* Ensure detailed documentation exists for the data source, formats, and usage patterns.
* • Are a data model and data catalog available?
* A clear schema (data model) and a searchable inventory of data (data catalog) streamline understanding and integration.

## 2. Extract & Load (ETL/ELT Processes)

Define how data will be moved into the warehouse:

* • Incremental vs. Full Loads?
* Determine whether the process involves loading all the data (full) or only changes since the last load (incremental).
* • Data Scope & Historical Needs:
* Clarify the range of data to extract, such as past 5 years, current data, or specific subsets.
* • What is the expected size of the extracts?
* Plan for the volume of data being moved to avoid pipeline bottlenecks.
* • Are there any data volume limitations?
* Confirm system or infrastructure constraints that could restrict the data transfer.
* • How do we avoid impacting the source system’s performance?
* Strategies like limiting extract frequency or scheduling during off-peak hours can prevent disruptions.
* • What authentication and authorization methods are required?
* Implement secure access mechanisms (e.g., tokens, SSH keys, VPN, or IP whitelisting).

## 3. Architecture & Technology Stack

Evaluate the infrastructure and tools to support integration:

* • How is the data currently stored?
* Identify the platforms or storage solutions (SQL Server, Oracle, AWS, Azure, Google BigQuery, etc.) used.
* • What are the integration capabilities?
* Confirm supported methods like APIs, Kafka, file-based extracts, direct database connections, or middleware.

## 4. Data Quality & Governance

Ensure the reliability and consistency of data:

* • Is data quality validated before ingestion?
* Identify checks for completeness, accuracy, and consistency during extraction.
* • Are there defined data governance policies?
* Set rules for data access, usage, and maintenance to ensure compliance with regulatory requirements.
* • How will sensitive data be handled?
* Include encryption, masking, or anonymization for personally identifiable or sensitive information.

## 5. Monitoring & Maintenance

Establish ongoing processes for system health and performance:

* • How will the ETL/ELT process be monitored?
* Use tools to track job completion, detect errors, and handle failures in real-time.
* • What is the plan for handling schema changes?
* Anticipate changes in source data structures and update pipelines accordingly.
* • Are there defined SLAs (Service Level Agreements)?
* Set expectations for data availability, latency, and accuracy to meet business needs.

## 6. Scalability & Future Growth

Ensure the system can handle increasing data volumes:

* • Is the warehouse scalable?
* Confirm whether storage and compute resources can scale with growing data or queries.
* • Are future data sources considered?
* Plan for integration of additional data sources as business needs evolve.