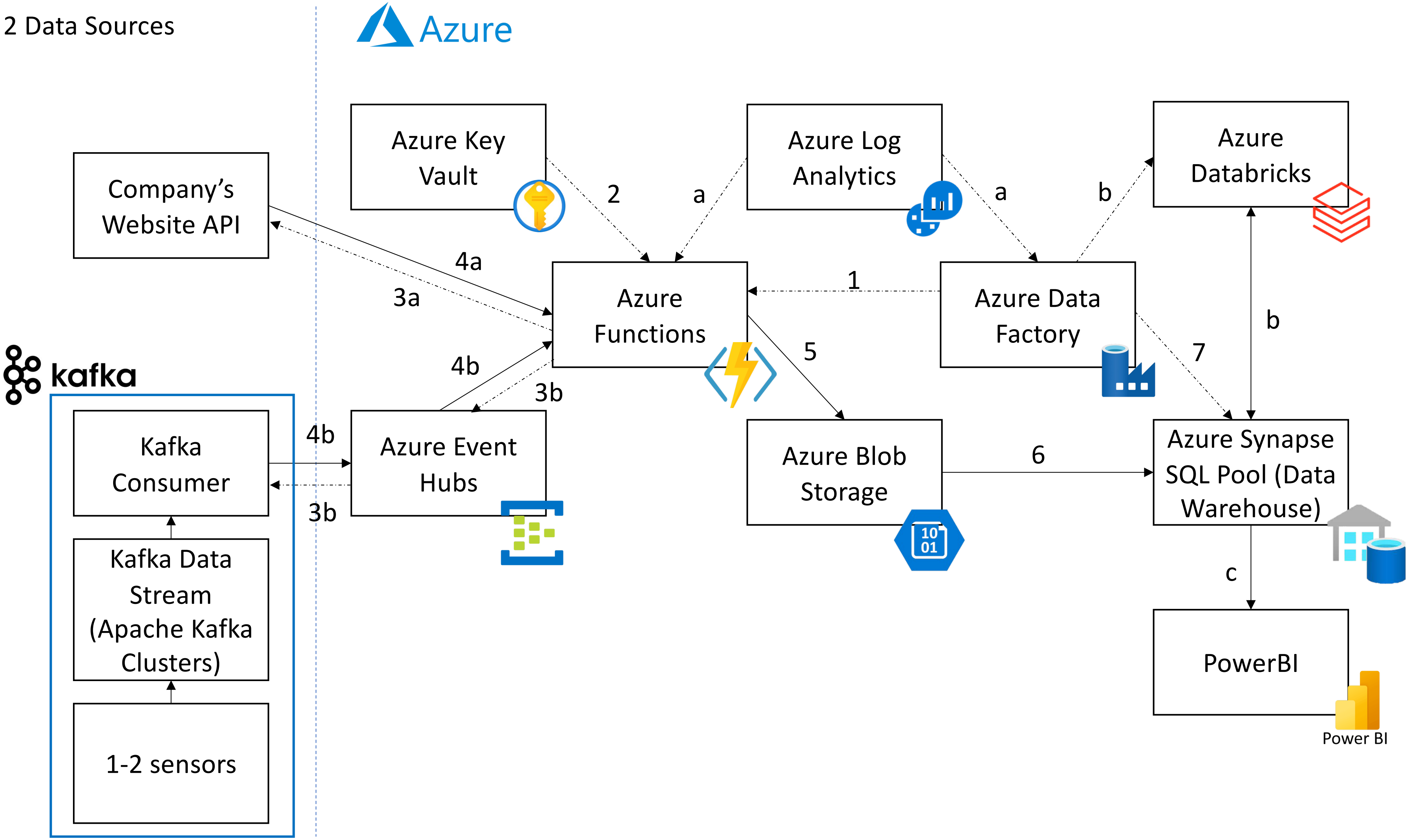

2 Data Sources

Azure

Company's Website API

Azure Key Vault

Azure Log Analytics

Azure Databricks

Azure Functions

Azure Data Factory

kafka

Kafka Consumer

Kafka Data Stream (Apache Kafka Clusters)

1-2 sensors

Azure Event Hubs

Azure Blob Storage

Azure Synapse SQL Pool (Data Warehouse)

PowerBI

Steps	Details
1	Azure Data Factory (ADF) provides a workflow-based data integration service by orchestrating data movement and data transformation activities. Pipelines and triggers can be authored in ADF and they consist of parameters like data source and tumbling window period (scheduled runs periodically). ADF is the main orchestrator of the data ingestion process as it calls upon and pass the required parameters to the ingestion framework (written by company’s software engineer and hosted in Azure Functions) when triggered.
2	The ingestion framework (hosted in Azure Functions) will first reads the secrets required to call the source system. The secrets, such as API keys, connection strings, or other sensitive information, are stored securely in Azure Key Vault. Hence, Azure Functions uses managed identities to authenticate and interact with Azure Key Vault without the need for explicit credentials. It will then retrieves the required secrets from Key Vault at runtime.
3a, 4a	The ingestion framework uses the secrets retrieved from Azure Key Vault to authenticate and call the Company API. The plugin associated with the source system handles this API call. The plugin then requests data from the source system for the specified tumbling window period. The data is fetched from the Company API and prepared for processing.
3b, 4b	Azure Event Hubs fetches data from Kafka consumers by providing a Kafka-compatible endpoint, allowing Kafka consumers to connect and pull data from Event Hubs directly. This Kafka endpoint in Azure Event Hubs enables seamless integration with existing Kafka applications and tools without any code changes. Azure Event Hubs can then easily connect to Kafka consumers to fetch and process the data, while ensuring data is distributed and processed efficiently, leveraging the scalability and durability of Azure Event Hubs.
5	The ingestion framework in Azure Functions writes the received and processed data to the main storage account, which is Azure Blob Storage. Azure Blob Storage provides a cost-effective and durable solution for storing large amounts of unstructured data. A data retention policy can also be specified to ensure that data is retained for the required 60-day period. Thereafter, Azure functions can be utilised to set up scheduled jobs or event-driven triggers to automatically purge these ‘expired’ data.
6	After data is stored in Azure Blob Storage, ADF can also calls stored procedures in Azure Synapse SQL Pool to transform and load data into the warehouse. The stored procedure can be designed to handle various aspects of data loading and processing, such as data validation, schema changes, or aggregations. End users such as data analysts can then work with cleaner datasets and generate better insights.
7	Due to the 60-day data retention policy, ADF can be used to run a stored procedure written in Azure Synapse SQl Pool to identify and delete any dataset that has ‘expired’. It can be scheduled to run automatically every 60 days through the trigger parameters in ADF.

Extras	Details
a	For Azure Functions, Log Analytics will enable real-time monitoring and analysis of function logs. It helps to empower engineers to identify performance issues, set up alerts, and respond proactively to incidents. This centralized logging solution provides a holistic view of function performance, ensuring high availability and rapid troubleshooting. For ADF, Log analytics provide better monitoring of data pipelines and workflows, enabling engineers to manage streams of data, identify failures, and improve productivity. By capturing relevant logs and telemetry data, engineers can quickly address issues and ensure data accuracy and compliance. Easy-to-configure and deploy Log Analytics simplifies analysis and frees up technical resources, resulting in more efficient and flexible data management for the entire system
b	Azure databricks can be used for limited set of data transformation (written in python) which can be difficult to do in SQL.
c	BI tool can runs automated daily extraction to read transformed data for the reporting data model.

Assumptions	Details
1	All data after 60 days will be deleted no matter if they are being used for analytical work etc.
2	Code written by company’s software engineer can be migrated and hosted in Azure seamlessly.
3	The stream application is hosted by the external company and our company only interact with Kafka consumer.