

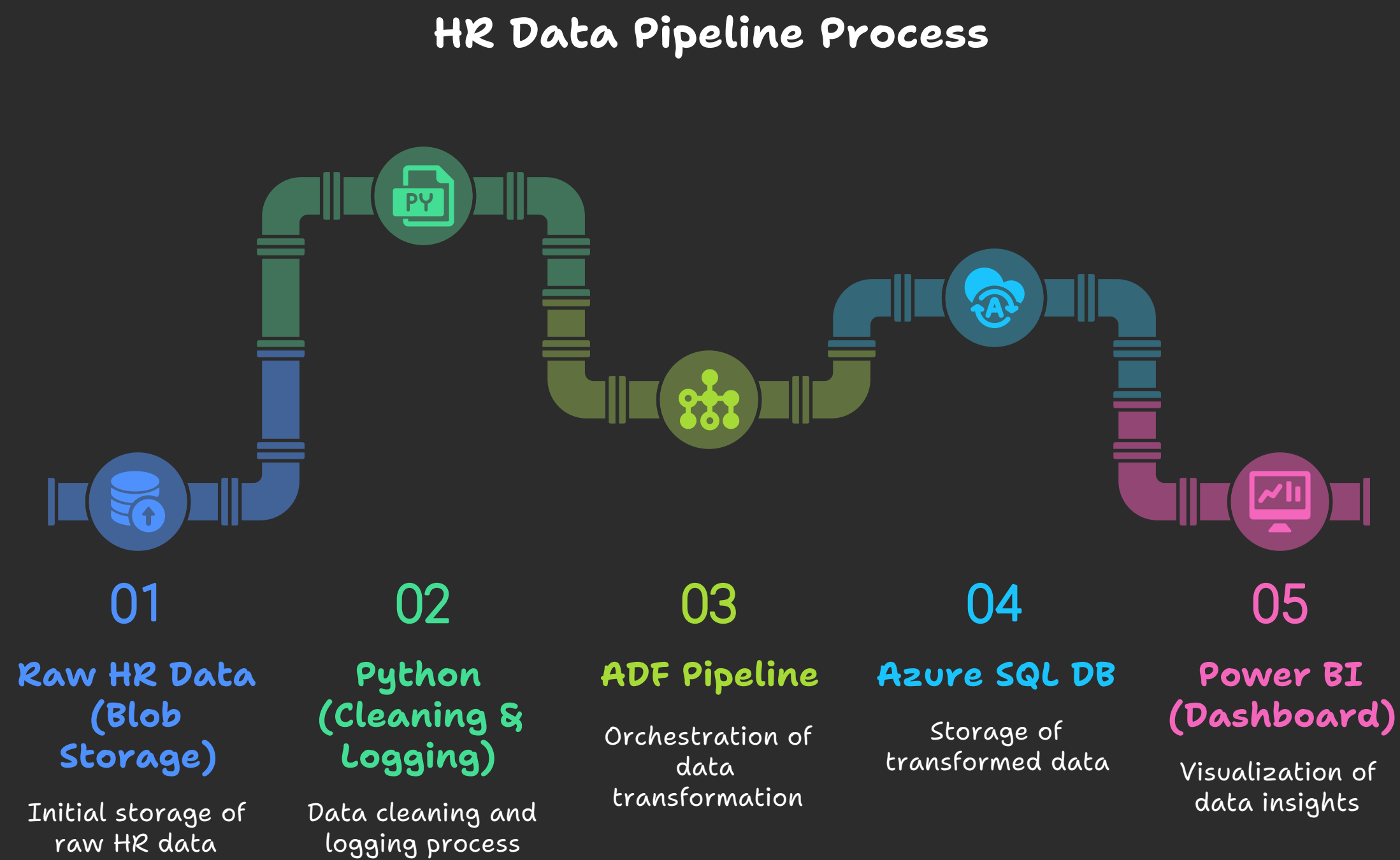


HR Data Pipeline: From Raw Data to Interactive Dashboard

This document outlines the flow of data in our Human Resources (HR) data pipeline, from its initial storage as raw data to its final presentation in an interactive Power BI dashboard. The pipeline encompasses data extraction, cleaning, transformation, loading, and visualization, leveraging various Azure services for efficiency and scalability. This document details each stage of the pipeline, highlighting the technologies used and their respective roles.

Pipeline Overview

The HR data pipeline follows a structured process to transform raw HR data into actionable insights. The data originates in Blob Storage, undergoes cleaning and logging via a Python script, is orchestrated by an Azure Data Factory (ADF) pipeline, is stored in an Azure SQL Database, and is finally visualized in a Power BI dashboard. The following diagram illustrates the data flow:



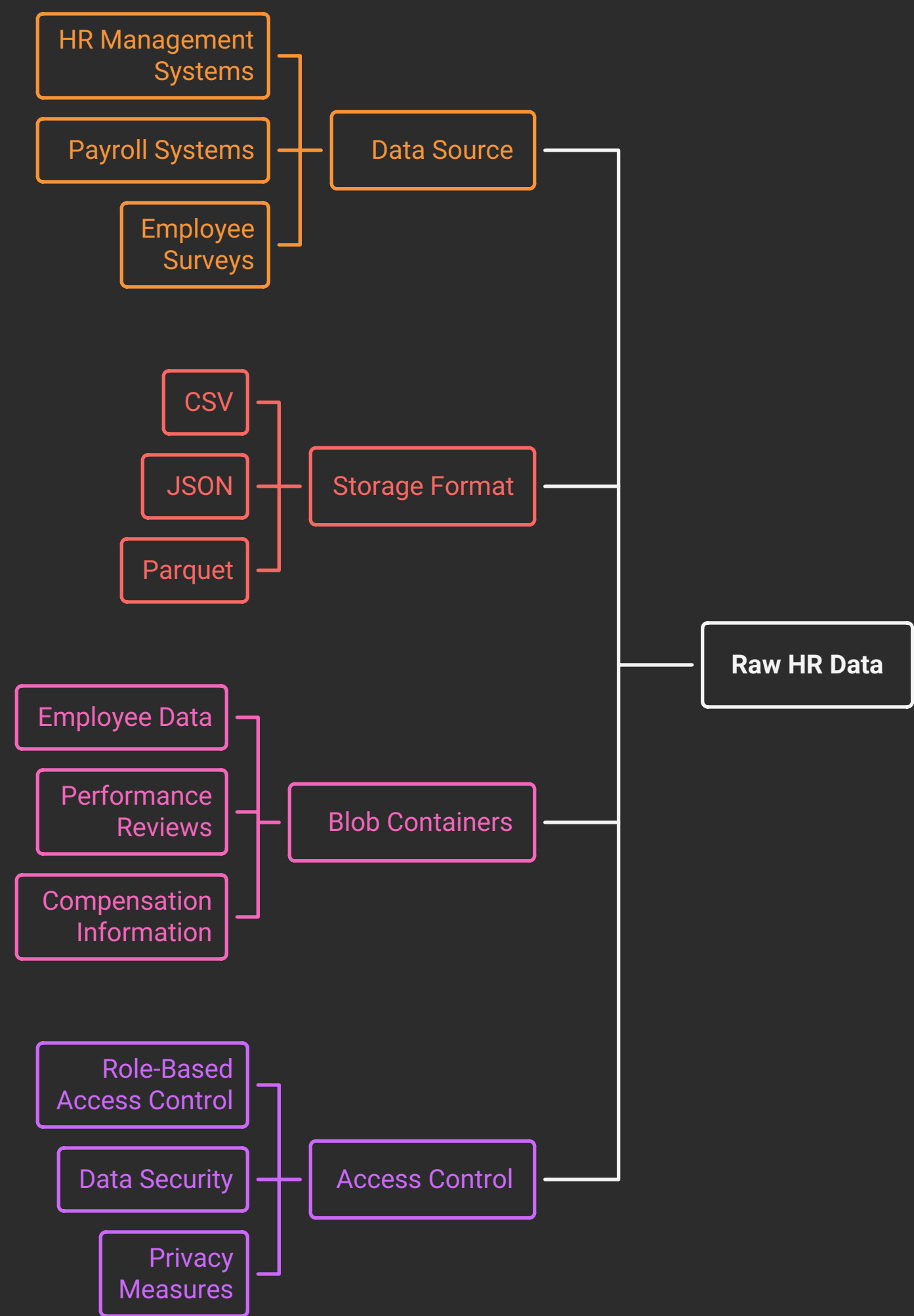
```
graph LR
    A[Raw HR Data (Blob Storage)] --> B(Python (Cleaning & Logging))
    B --> C{ADF Pipeline}
    C --> D[Azure SQL DB]
    D --> E(Power BI (Dashboard))
```

Stage 1: Raw HR Data (Blob Storage)

The initial stage involves storing the raw HR data in Azure Blob Storage. Blob Storage serves as a cost-effective and scalable repository for unstructured and semi-structured data.

- **Data Source:** The HR data originates from various sources, including HR management systems, payroll systems, and employee surveys.
- **Storage Format:** The data can be stored in various formats, such as CSV, JSON, or Parquet. The choice of format depends on the data structure and the requirements of subsequent processing stages.
- **Blob Containers:** Data is organized into blob containers for logical grouping and access control. For example, separate containers might be used for employee data, performance reviews, and compensation information.
- **Access Control:** Appropriate access controls are implemented to ensure data security and privacy. Role-Based Access Control (RBAC) is used to grant specific permissions to users and services.

HR Data Pipeline: Data Storage and Access Control



Made with  Napkin

Stage 2: Python (Cleaning & Logging)

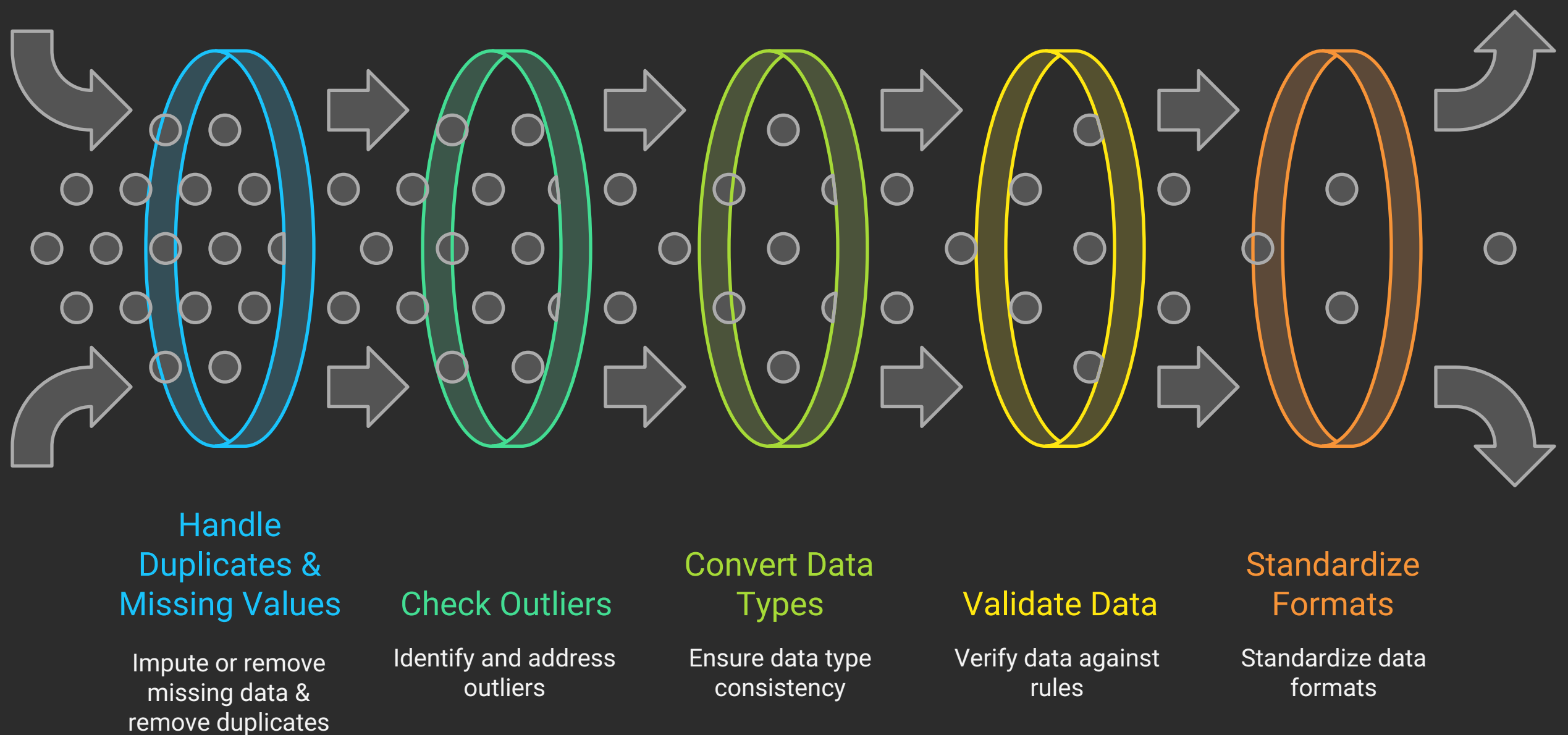
A Python script is used to clean and log the raw HR data before it is ingested into the Azure SQL Database. This stage is crucial for ensuring data quality and consistency.

Did Y data profiling for quick review

- **Data Cleaning:** The Python script performs several data cleaning tasks, including:
 - **Handling Missing & Duplicate Values:** Imputing or removing Duplicate & missing values based on the specific data field and business requirements.
 - **Checking for outliers:** checking outliers in age or any numerical formats.
 - **Data Type Conversion:** Converting data types to ensure consistency and compatibility with the Azure SQL Database.
 - **Data Validation:** Validating data against predefined rules and constraints to identify and correct errors.
 - **Data Standardization:** Standardizing data formats, such as date formats and address formats.
- **Logging:** The Python script logs all data cleaning activities, including the number of records processed, the number of errors encountered, and the actions taken to correct the errors. This logging information is valuable for auditing and troubleshooting purposes.

- **Libraries:** The Python script utilizes libraries such as Pandas for data manipulation, NumPy for numerical operations.

HR Data Cleaning Process








Made with  Napkin

Stage 3: ADF Pipeline

Azure Data Factory (ADF) is used to orchestrate the data pipeline, automating the movement and transformation of data from Blob Storage to Azure SQL Database.

- **Pipeline Activities:** The ADF pipeline consists of several activities, including:
 - **Copy Activity:** Copies the cleaned HR data from Blob Storage to Azure SQL Database.
 - **Data Transformation Activity:** Executes stored procedures in Azure SQL Database to further transform the data.
 - **Lookup Activity:** Retrieves metadata from Azure SQL Database to control the flow of the pipeline.
- **Triggers:** The ADF pipeline can be triggered on a schedule, on demand, or in response to events. Scheduled triggers are used to run the pipeline at regular intervals, while event-based triggers are used to run the pipeline when new data is available in Blob Storage.
- **Linked Services:** ADF uses linked services to connect to various data sources and compute resources. Linked services are configured for Blob Storage and Azure SQL Database.
- **Datasets:** Datasets define the structure and location of the data used in the pipeline. Datasets are configured for the HR data in Blob Storage and the tables in Azure SQL Database.
- **Monitoring:** ADF provides built-in monitoring capabilities to track the progress of the pipeline and identify any errors.

ADF Pipeline Components

Component	Description
 Activities	Copy Activity: Copies data from Blob Storage to Azure SQL Database.
 Triggers	Scheduled triggers: Run the pipeline at regular intervals.
 Linked Services	Connects to Blob Storage and Azure SQL Database.
 Datasets	Defines the structure and location of the HR data in Blob Storage and the tables in Azure SQL Database.
 Monitoring	Tracks the progress of the pipeline and identifies any errors.

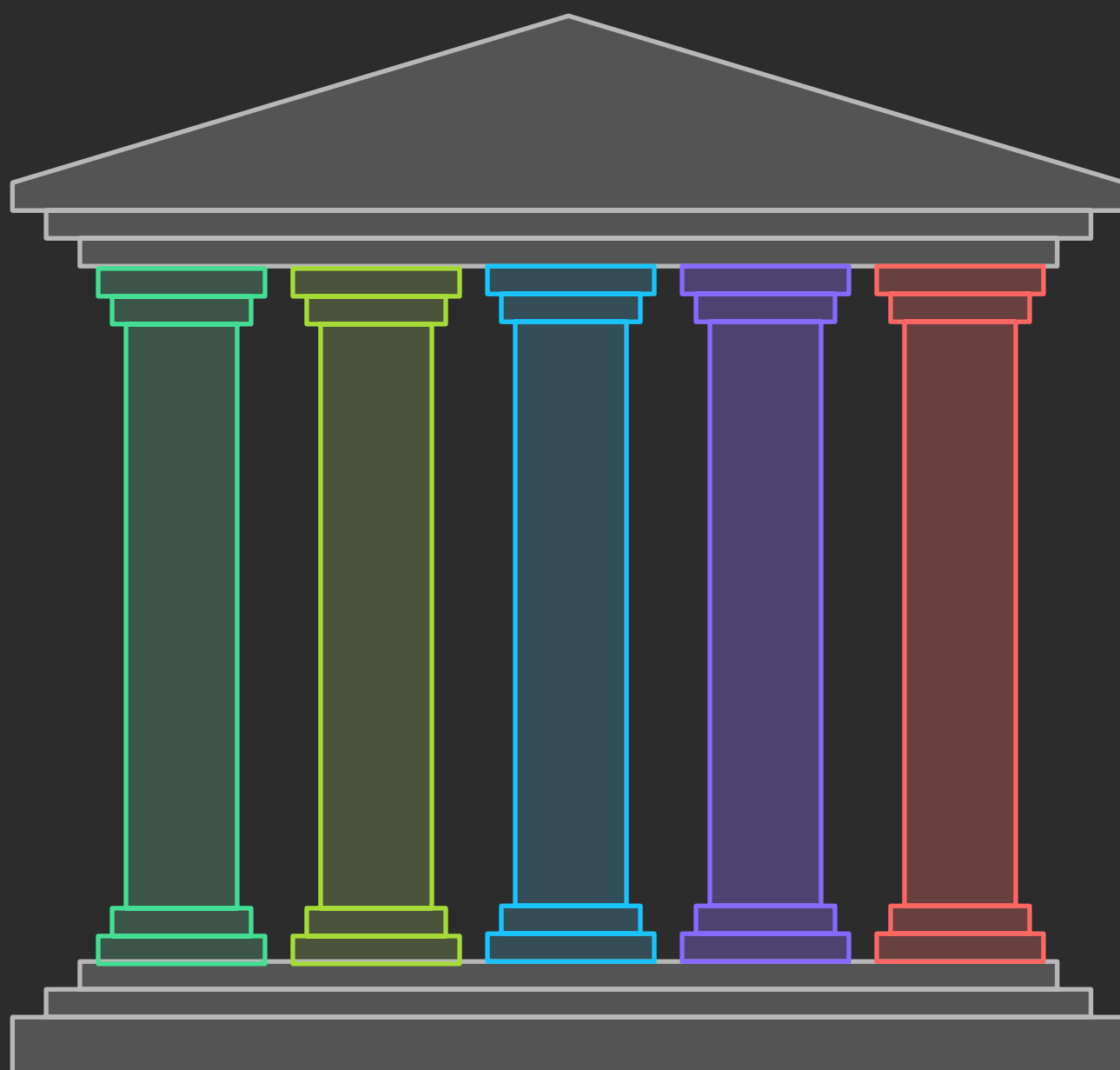
Made with  Napkin

Stage 4: Azure SQL DB

Azure SQL Database serves as the relational database for storing the transformed HR data.

- **Database Schema:** A well-defined database schema is created to store the HR data in a structured manner. The schema includes tables for employee data, performance reviews, compensation information, and other relevant data.
- **Data Types:** Appropriate data types are used for each column in the tables to ensure data integrity and consistency.
- **Indexes:** Indexes are created on frequently queried columns to improve query performance.
- **Stored Procedures:** Stored procedures are used to perform data transformations and calculations within the database.
- **Security:** Security measures are implemented to protect the data in Azure SQL Database. These measures include:
 - **Firewall Rules:** Configuring firewall rules to restrict access to the database.
 - **Authentication:** Using Azure Active Directory [Azure AD] authentication to manage user access.

Azure SQL Database Structure



Database Schema

A structured framework for organizing HR data tables.



Data Types

Ensuring data integrity through appropriate column types.



Indexes

Enhancing query performance with indexed columns.



Stored Procedures

Performing data transformations and calculations efficiently.



Security

Protecting data with firewall rules and authentication.

Made with  Napkin

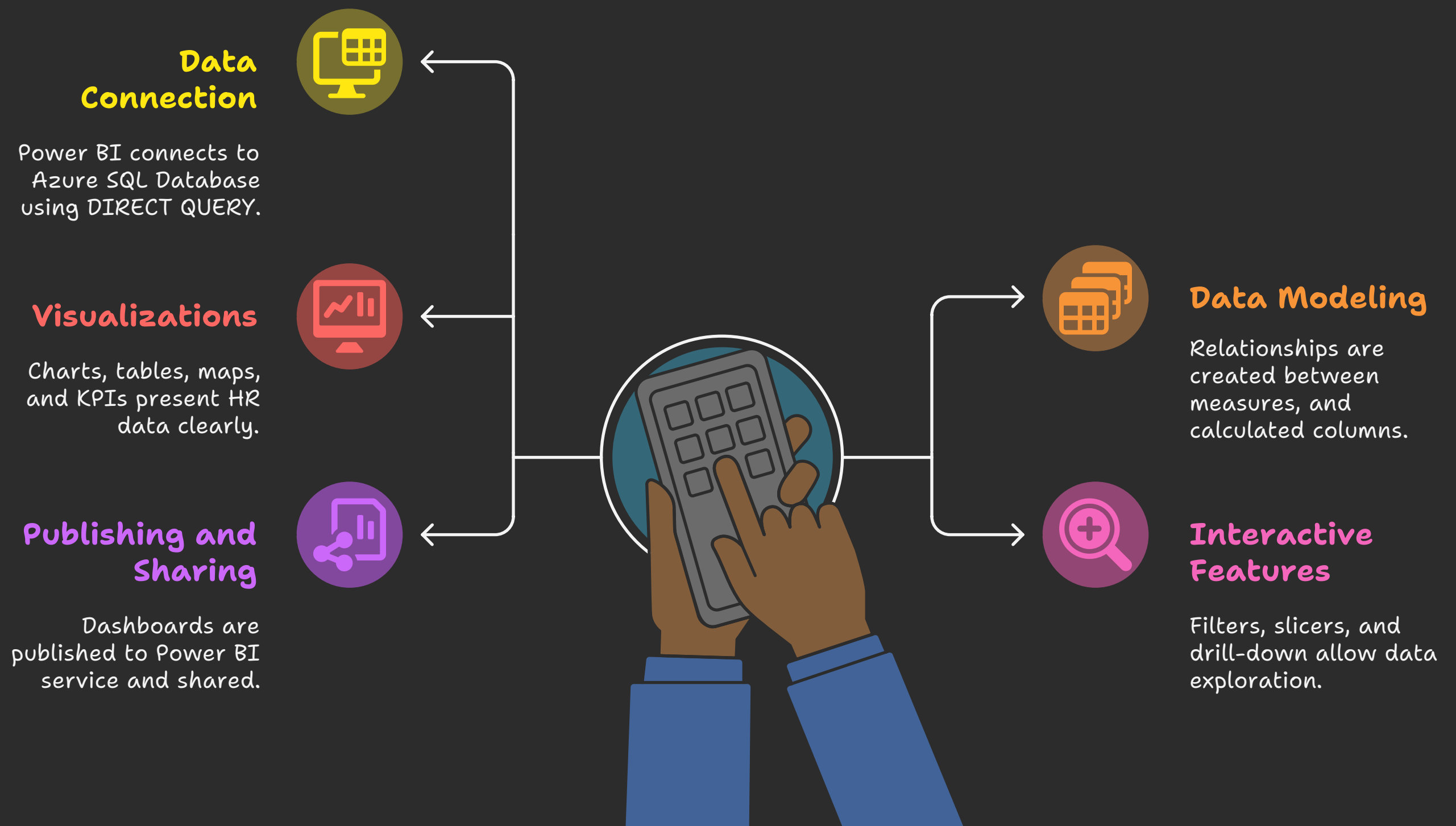
Stage 5: Power BI (Dashboard)

Power BI is used to create interactive dashboards that visualize the HR data and provide actionable insights.

- **Data Connection:** Power BI connects to the Azure SQL Database to retrieve the transformed HR data through **DIRECT QUERY**.
- **Data Modeling:** Data modeling techniques are used to create relationships between the tables in the database and to define measures and calculated columns.
- **Visualizations:** Various visualizations are used to present the HR data in a clear and concise manner. These visualizations include:
 - **Charts:** Bar charts, line charts, pie charts, and scatter plots are used to visualize trends and patterns in the data.
 - **Tables:** Tables are used to display detailed data.
 - **Maps:** Maps are used to visualize geographic data.
 - **Key Performance Indicators (KPIs):** KPIs are used to track progress towards specific goals.
- **Interactive Features:** Power BI dashboards provide interactive features that allow users to explore the data and drill down into specific details. These features include:
 - **Filters:** Filters are used to narrow down the data displayed in the dashboard.
 - **Slicers:** Slicers are used to filter the data based on specific values.
 - **Drill-Down:** Drill-down allows users to navigate from a high-level summary to more detailed data.

- **Publishing and Sharing:** Power BI dashboards can be published to the Power BI service and shared with other users.

Power BI Dashboard Features



Made with  Napkin

Conclusion

This HR data pipeline provides a comprehensive solution for transforming raw HR data into actionable insights. By leveraging Azure services such as Blob Storage, Azure Data Factory, Azure SQL Database, and Power BI, the pipeline ensures data quality, scalability, and security. The interactive Power BI dashboards empower HR professionals to make data-driven decisions and improve workforce management.

HR Data Pipeline Process

