**EXTRACT**

───────────────────────────────────────────────────────

Step #1: Fake Data Generation & CSV Export

• Environment: Jupyter Notebook

• Tools: Faker, Pandas

• Actions: Synthesize job market data (job\_title, industry, salary,

years\_of\_experience, ai\_replacement\_risk) and export as a CSV file

───────────────────────────────────────────────────────

↓

───────────────────────────────────────────────────────

Step #2: Configure Azure PostgreSQL Flexible Server & Connect to

Azure Data Studio

• Environments: Azure Portal, Azure Data Studio

• Actions: Provision the server, choose region/compute, set admin

credentials, enable public access, configure firewall rules, retrieve

host/port/database name/username, and verify connection

───────────────────────────────────────────────────────

**TRANSFORM**

───────────────────────────────────────────────────────

Step #3: Schema & Table Creation (SQL in ADS)

• Actions: Create a new schema and target table with columns matching

the CSV structure (e.g., job\_title, salary, ai\_replacement\_risk),

using appropriate data types (VARCHAR, INT, SERIAL) and constraints

───────────────────────────────────────────────────────

↓

───────────────────────────────────────────────────────

Step #4: Data Cleaning & Transformation (Python & Pandas in ADS)

• Actions:

- Load CSV data into a Pandas DataFrame

- Clean the data: remove underscores, rename columns, convert string and

numeric fields to appropriate data types, remove nulls/unrealistic records

- Add calculated columns (salary\_group, experience\_level, work\_type)

───────────────────────────────────────────────────────

**LOAD**

───────────────────────────────────────────────────────

Step #5: Load to Azure PostgreSQL Cloud Database (Python & SQLAlchemy in ADS)

• Actions:

- Create an engine using SQLAlchemy

- Load the cleaned and transformed data (using DataFrame.to\_sql())

- Verify table population and record integrity

───────────────────────────────────────────────────────

↓

───────────────────────────────────────────────────────

Step #6: Visualizations & Insights (Tableau Public)

• Actions:

- Connect Tableau to the Azure PostgreSQL server

- Build interactive dashboards (e.g., average AI replacement risk by industry,

filters for job title, experience level, work type; gradient-colored bar charts)

───────────────────────────────────────────────────────

**Tech Stack**

**Languages & Libraries**

* **SQL –** Schema and Table creation
* **Python** – Core scripting for data generation, transformation, and database loading
* **Pandas** – Data manipulation, cleaning, and transformation
* **Faker** – Synthetic data generation
* **SQLAlchemy** – Database connectivity for PostgreSQL DB
* **psycopg2** – PostgreSQL database driver for Python

**Database & Cloud**

* **Azure PostgreSQL Flexible Server** – Cloud-hosted relational database for storing the transformed data
* **SQL (PostgreSQL)** – Table schema creation, data definition, and querying
* **Azure Data Studio (ADS)** – SQL development and Python integration environment for managing and interacting with PostgreSQL on Azure

**Development & Analysis Tools**

* **Jupyter Notebook** – Python-based environment, used for data generation and initial CSV export)
* **Azure Data Studio** – Used for SQL development, schema setup, data cleaning, and Python execution
* **Tableau Public** – Data visualization platform used for visualization and dashboard creation leading to insights