* **Title**: Data Scraping, Processing, and Visualization Using Azure,GitHub , PowerBI &Tableau
* **Prepared by**:  
  Ahmed Abdelrahan  
  Abedrhman Hany  
  Omar Elmnofy  
  Mohammed Baligh
* **Supervised by**: Eng Khalid Seif
* **University Name**
* **Graduation Date**

**Table of Contents (Page 2)**

* Introduction
* Data Scraping Overview
* Scraping Workflow
* Scraping Challenges
* Azure Function for Data Upload
* Pipeline 1 – Raw Data Ingestion
* Pipeline 2 – Data Transformation and Warehousing
* Data Modeling
* PowerBI Dashboard Overview
* PowerBI Visuals
* Key Insights
* Conclusion
* Future Work
* Challenges and Solutions
* Team Credits

**1. Introduction (Page 3)**

* **Objective**: A comprehensive project focused on scraping data from Amazon’s regional websites (Egypt, India, Saudi Arabia, Japan), processing it using Azure Functions, and building pipelines for data transformation, warehousing, and visualization using PowerBI.
* **Motivation**: Explain why e-commerce data analysis is important and how cross-regional insights provide a competitive advantage.
* **Technologies Used**: Python (for scraping), Azure Functions, Data Pipelines, PowerBI.

**2. Data Scraping Overview (Page 4)**

* **Regions Covered**: Amazon Egypt, India, Saudi Arabia, Japan.
* **Data Collected**: Product name, price, availability, ratings, product category.
* **Scraping Tools**: Python libraries used:
* **BeautifulSoup** for parsing HTML.
* **Scrapy** for crawling and scraping.
* **Selenium** for handling JavaScript-loaded content.

**3. Scraping Workflow (Page 5)**

* **Step-by-Step Process**:
* Sending requests to Amazon websites.
* Parsing HTML and extracting data.
* Cleaning and normalizing extracted data.
* Storing the data in CSV or JSON format.
* **Automation**: Discuss how Python scripts automate scraping, handle exceptions, and retry failed requests.

**4. Scraping Challenges (Page 6)**

* **Key Challenges**:
* **CAPTCHAs**: Managed using proxies and delay handling.
* **IP Blocking**: Solutions like rotating IP addresses.
* **Dynamic Content**: Handled via Selenium for JavaScript-rendered pages.
* **Ethical Considerations**: Addressing Amazon’s scraping policies and ethical scraping practices.

**5. Azure Function for Data Upload (Page 7)**

* **Azure Function Overview**: Explain how serverless Azure Functions process scraped data.
* **Steps**:
* Receiving scraped data as JSON.
* Preprocessing and validating data.
* Storing data in **Azure Blob Storage** (Raw Data folder).
* **Scalability**: Benefits of using Azure Functions to automatically handle large-scale data uploads.

**6. Pipeline 1 – Raw Data Ingestion (Page 8)**

* **Pipeline Overview**: The first pipeline takes raw scraped data and ingests it into a **Raw Folder** in Azure.
* **Process**:
* Data ingestion from Azure Function.
* Storing and organizing raw data based on region and timestamp.
* **Tools**: Azure Data Factory for pipeline orchestration.

**7. Pipeline 2 – Data Transformation and Warehousing (Page 9)**

* **Data Transformation**:
* Cleaning data (removing duplicates, handling missing values).
* Normalizing currencies and product categories across regions.
* **Loading to Warehouse**: Transformed data loaded into an **Azure SQL Data Warehouse** for further analysis.
* **Azure Data Factory**: Automating data flow and transformation.

**8. Data Modeling (Page 10)**

* **Star Schema Design**:
* **Fact Table**: Storing metrics (e.g., sales data, prices).
* **Dimension Tables**: Product details, regions, time-based information.
* **Why Star Schema?**: Efficient querying and easy-to-use structure for PowerBI reports.

**9. PowerBI Dashboard Overview (Page 11)**

* **Connecting PowerBI to Data Warehouse**: Steps for setting up PowerBI with the data warehouse.
* **Dashboard Layout**: Overview of dashboard features:
* Filters for region, time, and category.
* Interactive visual elements (graphs, charts).

**10. PowerBI Visuals (Page 12)**

* **Key Visualizations**:
* **Price Trends**: Line graphs showing price variations over time across regions.
* **Top Products**: Bar charts showing the most popular products in each region.
* **Regional Comparison**: Maps visualizing product availability and pricing in different regions.

**11. Key Insights (Page 13)**

* **Insights Gained**:
* **Regional Price Variation**: Notable price differences for the same product across different regions.
* **Best-Selling Products**: Insights into product popularity based on region.
* **Seasonality**: Seasonal fluctuations in prices and availability during holidays or sales events.

**12. Conclusion (Page 14)**

* **Project Summary**:
* Successfully built a multi-regional data scraping pipeline.
* Automated data ingestion, transformation, and storage using Azure services.
* Visualized key insights using PowerBI, providing valuable e-commerce intelligence.
* **Key Takeaways**: Discuss the technical and business advantages provided by the project.

**13. Future Work (Page 15)**

* **Suggestions for Expansion**:
* **Additional Regions**: Scraping Amazon websites from other countries (e.g., US, UK).
* **Real-Time Data**: Automating real-time scraping and pipeline execution for continuous updates.
* **Machine Learning Integration**: Predictive analytics to forecast pricing trends or demand spikes.

**14. Challenges and Solutions (Page 16)**

* **Challenges Encountered**:
* Handling large volumes of data.
* Ensuring data consistency across regions.
* **Solutions**:
* Using parallel processing for faster scraping and transformations.
* Implementing data validation scripts to ensure quality.

**15. Team Credits (Page 17)**

* **Team Members**:
* Ahmed Abdelrahan
* Abedrhman Hany
* Omar Elmnofy
* Mohammed Baligh
* **Supervised by**: Eng Khalid Seif

**References (Page 18)**

* **Citations**: List of tools, documentation, libraries, and research used during the project.