

DATA ENGINEERING APPROACH TO GAIN INSIGHTS ON SEMICONDUCTOR CHIP PERFORMANCE

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

NIRMAL KUMAR P (Reg. No – 40110858)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

CATEGORY -1 UNIVERSITY BY UGC

Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHISALAI, CHENNAI - 600119

APRIL - 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Category - I University by UGC

Accredited "A++" by NAAC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Nirmal Kumar P (Reg.No - 40110858)** who carried out the Project entitled **"DATA ENGINEERING APPROACH TO GAIN INSIGHTS ON SEMICONDUCTOR CHIP PERFORMANCE"** under my supervision from January 2024 to April 2024.

Jany 15/4/24
Internal Guide

Dr. S.L. JANY SHABU M.TECH., Ph.D

[Signature]
Head of the Department

Dr. L. LAKSHMANAN, M.E., Ph.D.

HEAD OF THE DEPARTMENT
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Jeyaraj Nagar, Rajiv Gandhi Salai,
Chennai - 600 119.

Submitted for Viva voce

Examiner: Nirmal Kumar P

15.04.2024

[Signature]
15/4/24
Internal Examiner

[Signature]
15/4/24
External Examiner

DECLARATION

I, **NIRMAL KUMAR P (Reg.No- 40110858)**, hereby declare that the Project Report entitled **"DATA ENGINEERING APPROACH TO GAIN INSIGHTS ON SEMICONDUCTOR CHIP PERFORMANCE"** done by me under the guidance of **Dr. S. L. Jany Shabu, M.Tech., Ph.D** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE: 15.04.2024



PLACE :Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.**, Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. S. L. Jany Shabu, M.Tech., Ph.D**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-1 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

In the rapidly evolving landscape of semiconductor technology, the performance optimization of semiconductor chips is of paramount importance. This project delves into the realm of data engineering to unlock valuable insights into semiconductor chip performance. As semiconductor technology continues to push the boundaries of innovation, understanding and optimizing chip performance become crucial for enhancing overall computational efficiency.

The project has successfully navigated through the initial phases of data collection, encompassing a comprehensive dataset on semiconductor chip performance metrics. Initial data engineering processes, including data cleaning and preprocessing, have been executed to ensure the integrity and quality of the dataset. Ongoing efforts include the development of robust data pipelines for continuous data ingestion and integration from diverse sources. Additionally, exploratory data analysis techniques have been employed to identify patterns and trends in the semiconductor chip performance data.

Moving forward, the project aims to implement advanced machine learning algorithms and statistical models to extract meaningful insights from the accumulated data. Feature engineering will be a focal point, enhancing the predictive capabilities of the models. Furthermore, the project intends to explore anomaly detection techniques to identify irregularities in chip performance, contributing to the development of proactive maintenance strategies. The integration of real-time monitoring and feedback loops is also proposed to create a dynamic system capable of adapting to changing performance requirements.

The overarching goal is to establish a data-driven framework that not only provides a deep understanding of semiconductor chip performance but also enables proactive measures for optimization. The project envisions contributing to the field of semiconductor technology by providing actionable insights for both chip designers and manufacturers, ultimately fostering advancements in computational capabilities and efficiency.

Chapter No	TITLE		Page No.
	ABSTRACT		v
	LIST OF FIGURES		vii
1	INTRODUCTION		
	1.1	Project Focus	1
	1.2	Core Technological Framework	1
	1.3	Core Objectives	2
	1.4	Significance	2
2	LITERATURE SURVEY		
	2.1	Inferences from Literature Survey	3
	2.2	Open problems in Existing System	4
3	REQUIREMENTS ANALYSIS		
	3.1	Feasibility Studies/Risk Analysis of the Project	5
	3.2	Software Requirements Specification Document	6
4	DESCRIPTION OF PROPOSED SYSTEM		
	4.1	Selected Methodology or process model	
	4.2	Architecture / Overall Design of Proposed System	8
	4.3	Description of Software for Implementation and Testing plan of the Proposed Model/System	9
	4.4	Project Management Plan	11
	4.5	Project Plan Timeline	13
	4.6	Modules	14
	REFERENCES		15

LIST OF FIGURES

FIGURE NO	FIGURE NAME	Page No.
4.1	Methodology	15
4.2	System Architecture	16
4.3	Project Management Plan	20

CHAPTER 1

INTRODUCTION

In the relentless pursuit of advancing semiconductor technology, our project, titled "Data Engineering Approach to Gain Insights on Semiconductor Chip Performance," stands at the intersection of innovation, precision, and data-driven excellence. This initiative is conceived with the overarching goal of comprehensively assessing and optimizing semiconductor chip performance, employing a strategic amalgamation of cutting-edge tools and methodologies.

1.1 Project Focus:

The crux of our project lies in the meticulous integration of Microsoft Azure's Fabric, Python, and web scraping techniques. This triumvirate of technologies forms the backbone of our data engineering approach, enabling a nuanced exploration of semiconductor chip performance across a spectrum of performance metrics.

1.2 Core Technological Framework:

Microsoft Azure's Fabric: Our choice of Microsoft Azure's Fabric ensures seamless and scalable data operations. The fabric's distributed computing capabilities are harnessed for efficient data processing and storage, ensuring the agility required for handling diverse and extensive datasets.

Python: The versatility of Python is leveraged for algorithmic implementations and data manipulation. Python's rich ecosystem of libraries and frameworks equips our project to handle complex mathematical operations, cryptographic tasks, and data analysis with unparalleled efficiency.

Web Scraping: The incorporation of web scraping techniques enhances our dataset's richness by extracting real-world data from diverse sources. This approach broadens the representativeness of our analyses, ensuring our evaluations mirror the complexity of scenarios encountered in practical applications.

1.3 Core Objectives:

At the heart of our endeavor is the meticulous construction of a robust data architecture. This architecture is designed to facilitate the secure handling of data in 2/3 Normal Form schemas, aligning with best practices in database design. Our primary objectives include:

Normalization (2/3NF): Structuring data to 2/3 Normal Form schemas to enhance efficiency, reduce redundancy, and fortify data integrity.

Security: Implementing stringent security measures to safeguard the confidentiality and integrity of the collected data, ensuring the trustworthiness of our evaluations.

1.4 Significance:

Beyond the immediate objectives, our project's significance lies in contributing to the evolution of semiconductor chip assessment methodologies. By marrying cutting-edge technologies with rigorous data engineering principles, we aim to redefine the parameters for evaluating chip performance, fostering advancements in computational efficiency, energy optimization, and technological innovation.

As we delve into the subsequent sections, we invite scrutiny into the intricacies of our methodologies, data architecture, and the synergistic interplay of technologies, promising a profound exploration into the heart of semiconductor chip performance.

CHAPTER 2

LITERATURE SURVEY

2.1 INFERENCES FROM LITREATURE SURVEY

The literature survey reveals a landscape dominated by the imperative need for optimized data management and analytical processing, particularly in the context of multidimensional databases. These databases, specialized for data warehousing, data marts, and online analytical processing (OLAP) applications, excel in presenting high-level views of data, thereby facilitating the analysis of expansive record sets. The principal allure lies in their unparalleled query performance, enabling rapid retrieval of precise answers to user queries and fostering efficient data examination.

Recognizing the limitations of traditional data warehouses in addressing the escalating demands of contemporary enterprises, the paper advocates for a paradigm shift by proposing a novel method for designing data warehouse schemas. The focus centers on harnessing the capabilities of schema-free databases, commonly referred to as NoSQL databases, to integrate and analyze the diverse data emanating from social, mobile, and sensor sources. This pioneering approach aims to overcome the inherent constraints of conventional data warehousing systems.

Moreover, in the realm of cross-performance prediction for benchmark programs, the literature underlines the significance of more comprehensive technical elucidations, nuanced benchmark selection discussions, and a thorough exploration of potential limitations. The infusion of neural networks and regression analysis emerges as a transformative strategy, accommodating intricate non-linear relationships between input and output variables. This advanced methodology augments the precision of performance predictions, marking a notable advancement in the domain.

Particularly noteworthy is the introduction of tools such as Habitat and DNNPerf, each addressing critical needs in the realm of GPU selection and deep learning model performance prediction, respectively. Habitat distinguishes itself by aiding

users in making judicious decisions regarding GPU selection for DNN training, thereby exerting a considerable influence on overall training costs. On the other hand, DNNPerf fills a crucial void in the deep learning community by furnishing a tool for predicting the performance of deep learning models, thereby facilitating optimized model training and judicious resource allocation.

The recurrent theme of leveraging neural networks and regression analysis underscores a collective commitment towards accommodating the intricacies of non-linear relationships between variables. This unified approach is heralded as a pivotal stride forward, promising more accurate predictions in the dynamic landscape of performance analysis. In essence, the amalgamation of these insights contributes to a richer understanding of the evolving paradigms in database design, benchmarking, and predictive analytics, underscoring the interdisciplinary nature and the transformative potential of contemporary methodologies.

2.2 OPEN PROBLEMS IN EXISTING SYSTEM

The existing system faces several open problems that warrant attention and resolution. One notable challenge lies in the scalability of the data architecture, particularly as the volume of semiconductor chip performance data continues to grow exponentially. The potential bottlenecks in data processing and analysis pose obstacles to efficient system performance. Moreover, concerns related to data security and privacy persist, demanding ongoing efforts to fortify the system against evolving threats. The integration of diverse datasets, especially from external sources through web scraping, introduces complexities related to data consistency and accuracy, requiring meticulous validation and cleansing procedures. Additionally, there is a need for continual optimization of algorithms, as the computational demands for performance analysis and machine learning models may face scalability issues over time. Addressing these open problems is crucial for ensuring the reliability, security, and effectiveness of the existing system in the rapidly evolving landscape of semiconductor chip technology.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 FEASIBILITY STUDIES/RISK ANALYSIS OF THE PROJECT

Data Security and Privacy Concerns:

Risk: The project involves handling sensitive data related to semiconductor chip performance, raising concerns about data security and privacy.

Mitigation: Implement robust encryption protocols, access controls, and adhere to data protection regulations to ensure the confidentiality and integrity of the dataset.

Technical Challenges in Data Integration:

Risk: Integrating diverse datasets from various sources may pose technical challenges, leading to inconsistencies and inaccuracies.

Mitigation: Conduct thorough data profiling and cleansing, establish clear data integration protocols, and utilize advanced data engineering tools to ensure data quality.

Algorithmic Complexity:

Risk: Implementing complex algorithms in Python for performance analysis may lead to computational challenges and potential bottlenecks.

Mitigation: Conduct extensive testing of algorithms, optimize code for efficiency, and consider parallel processing techniques to handle computational loads.

Dependency on External Data Sources:

Risk: Relying on external data sources for web scraping introduces a dependency that may be subject to changes or disruptions.

Mitigation: Regularly monitor and validate external data sources, implement fallback mechanisms, and have contingency plans for disruptions.

Scalability Challenges:

Risk: As the project progresses, the scalability of the data architecture may become a concern, especially if the dataset grows exponentially.

Mitigation: Design the architecture with scalability in mind, leverage cloud-based services for dynamic scaling, and continuously monitor performance metrics.

3.2 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

Python:

Python, a versatile and powerful programming language, serves as the primary coding language for the project. Its extensive libraries and frameworks facilitate diverse functionalities, from data processing to algorithm implementation. Python's readability and flexibility make it an ideal choice for the data engineering approach, enabling efficient development and integration of algorithms.

Microsoft Fabric:

Microsoft Fabric, a robust microservices and distributed systems platform, forms a critical component of the project's architecture. It provides the infrastructure for scalable and resilient application development, allowing seamless deployment, management, and scaling of microservices. Microsoft Fabric enhances the overall system reliability and efficiency through its advanced features and integration capabilities.

BeautifulSoup library:

BeautifulSoup, a Python library, is instrumental in web scraping tasks within the data engineering approach. It simplifies the extraction of data from HTML and XML documents, making it an essential tool for gathering information from diverse online sources. The library's flexibility and ease of use contribute to effective data acquisition, a key aspect of gaining insights into semiconductor chip performance.

Git:

Git, a distributed version control system, plays a pivotal role in managing and tracking changes in the project's source code. It ensures collaboration among team members by maintaining a coherent and versioned codebase.

Microsoft Power App:

Microsoft Power App provides a user-friendly platform for developing custom applications without extensive coding. In the context of the project, Power App could be utilized to create intuitive interfaces or dashboards for specific user interactions or data visualization.

Microsoft Power BI:

Microsoft Power BI is a business analytics tool crucial for transforming raw data into meaningful insights. It offers powerful data visualization and reporting capabilities, enabling the creation of interactive dashboards and reports. Power BI integrates seamlessly with various data sources, making it an integral part of the project for presenting and interpreting semiconductor chip performance data effectively.

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

The proposed system, adopting a Medallion Architecture, represents an avant-garde Data Engineering Approach meticulously designed to unravel nuanced insights into Semiconductor Chip Performance. This innovative framework seamlessly integrates diverse data sources, employing advanced algorithms, and leveraging the power of Medallion Architecture to ensure efficient handling of data in 2/3NF schemas.

Through intricate layers of processing, the system aims to enhance the analysis and comprehension of semiconductor chip performance, ultimately contributing to the optimization of chip design and functionality.

The Medallion Architecture acts as the cornerstone, fostering a secure and structured environment for the exploration and extraction of valuable performance metrics from semiconductor chip data. This proposed system stands as a beacon in the realm of semiconductor engineering, promising a holistic understanding and optimization of chip performance through the lens of cutting-edge data engineering methodologies.

4.1 SELECTED METHODOLOGY OR PROCESS MODEL

Phase	Tasks
1. Project Initiation and Planning:	- Define project objectives and scope. - Identify key stakeholders. - Select technologies (Microsoft Azure, Python). - Allocate resources and plan project timelines.
2. Data Collection and Processing:	- Identify relevant data sources for chip performance. - Implement web scraping for real-world data. - Ingest data into Microsoft Azure Fabric for distributed processing.
3. Algorithm Development and Testing:	- Develop algorithms in Python (NumPy, Pandas, Scikit-learn). - Implement machine learning models for performance analysis. - Conduct unit testing with pytest for code integrity.
4. Database Design and Security Implementation:	- Design database schema adhering to 2/3NF normalization. - Implement normalization in SQL Server. - Implement robust security measures (encryption, access controls).
5. Web Scraping Enhancement and Monitoring:	- Continuously monitor web-scraped data sources. - Implement fallback mechanisms for disruptions.
6. Documentation and Reporting:	- Use Jupyter Notebooks for code documentation. - Generate comprehensive reports on chip performance insights.
7. Deployment and Optimization:	- Containerize applications using Docker for consistency. - Deploy on Microsoft Azure Cloud Services for scalability. - Implement real-time monitoring and iterative optimization.
8. Review and Iteration:	- Evaluate project outcomes against objectives. - Identify areas for improvement. - Iterate on methodologies for enhanced results.

Fig 4.1: Methodology

4.2 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

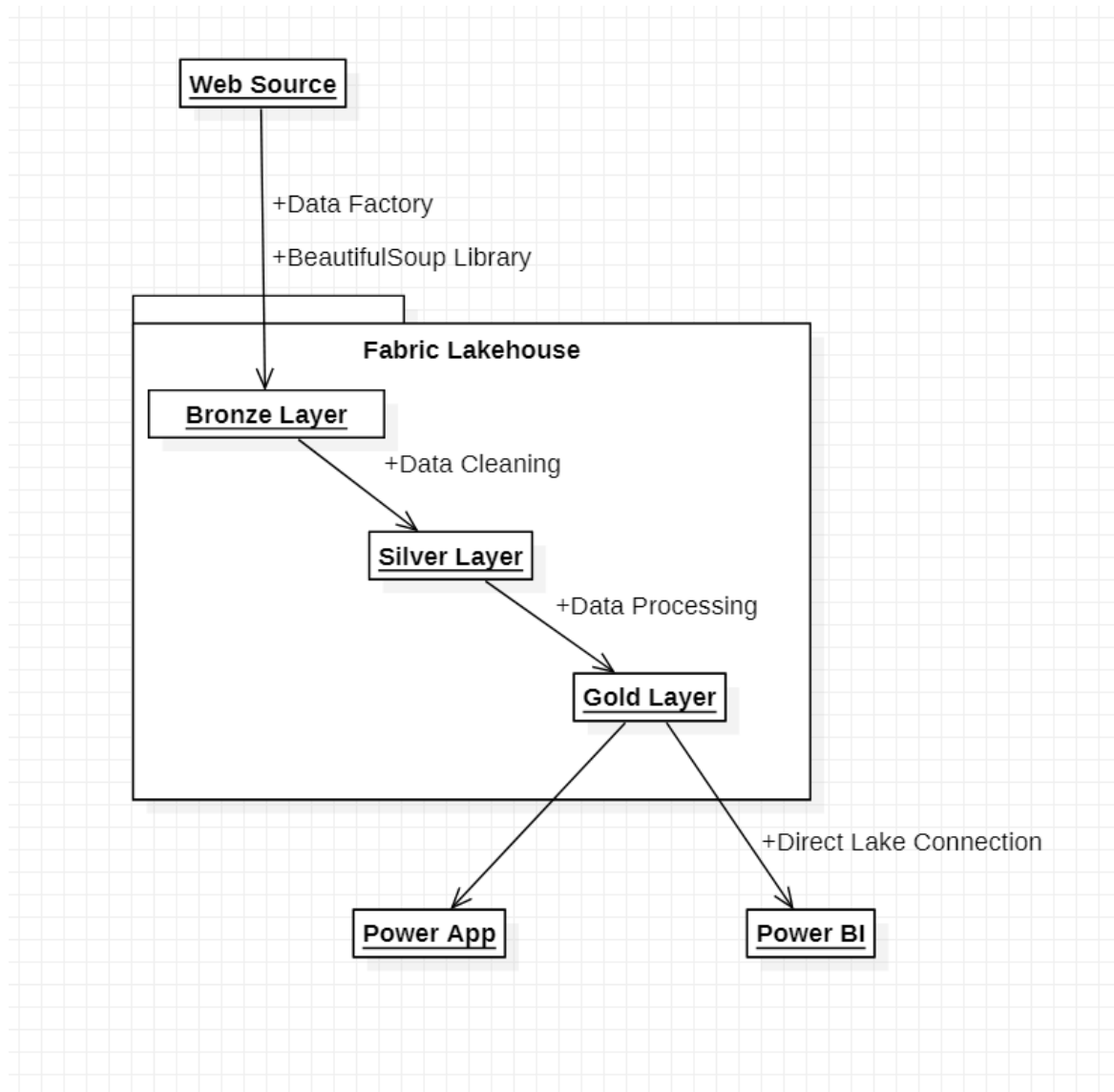


Fig 4.2: System Architecture

1. Overview:

The proposed system architecture aims to provide a robust and scalable framework for the project, "Data Engineering Approach to Gain Insights on Semiconductor Chip Performance." This architecture seamlessly integrates data engineering tools, web scraping, and advanced analytics within a secure and efficient ecosystem.

2. Components:

2.1 Python Ecosystem:

Core Programming Language: Python serves as the backbone for algorithm development, data manipulation, and analysis.

Libraries: NumPy, Pandas, and Scikit-learn for numerical operations, data manipulation, and machine learning algorithms.

2.2 Web Scraping Module:

Beautiful Soup: Employed for parsing HTML and XML data structures.

Selenium: Utilized for automating web browser interactions, ensuring efficient and reliable web scraping.

2.4 Database Management System:

SQL Server: Chosen as the DBMS for its robustness and support for normalized 2/3NF schemas.

2.5 Version Control:

Git: Implemented for version control, enabling collaborative development, tracking changes, and ensuring code integrity.

3. System Workflow:

3.1 Data Ingestion:

Raw data on semiconductor chip performance is ingested into Azure Fabric Services.

3.2 Data Processing:

Azure Fabric orchestrates distributed data processing, leveraging the power of Apache Spark for parallelized analytics.

3.3 Algorithm Implementation:

Python scripts within Jupyter Notebooks implement complex algorithms, leveraging Python libraries for numerical computations and machine learning.

3.4 Web Scraping:

Beautiful Soup and Selenium are employed to extract real-world data, enhancing the richness and diversity of the dataset.

3.5 Database Operations:

SQL Server manages structured data in 2/3NF schemas, ensuring data integrity and efficiency.

3.6 Security Measures:

Azure Security Services are integrated throughout the architecture to enforce encryption, access controls, and regulatory compliance.

4. Benefits:

Scalability: The architecture accommodates the dynamic nature of semiconductor performance data.

Flexibility: Components like Apache Spark and Python facilitate adaptive analyses and algorithm development.

Security: Robust security measures at multiple levels ensure the confidentiality and integrity of the data.

4.3 DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION AND TESTING PLAN OF THE PROPOSED MODEL/SYSTEM

1. Microsoft Fabric:

Description: Microsoft Azure Fabric is a versatile and comprehensive platform that plays a pivotal role in the seamless orchestration of distributed systems and microservices. It serves as a fundamental layer in building scalable, resilient, and efficient applications. Azure Fabric provides a set of tools and services for managing the complete application lifecycle, from development to deployment and monitoring.

Implementation: Utilize Azure Fabric Services for distributed computing and Azure Security Services for robust data security measures.

2. Python:

Description: Python is the core programming language chosen for algorithm development, data manipulation, and testing.

Implementation: Leverage Python libraries, including NumPy, Pandas, and Scikit-learn, for efficient numerical operations and machine learning implementations.

3. Web Scraping Tools:

Description: BeautifulSoup and Selenium are employed for web scraping to extract real-world data and enhance the diversity of the dataset.

Implementation: Develop scripts for automated web interactions and data extraction using BeautifulSoup and Selenium.

4. SQL Server:

Description: SQL Server is the chosen relational database management system (DBMS) for storing and managing structured data.

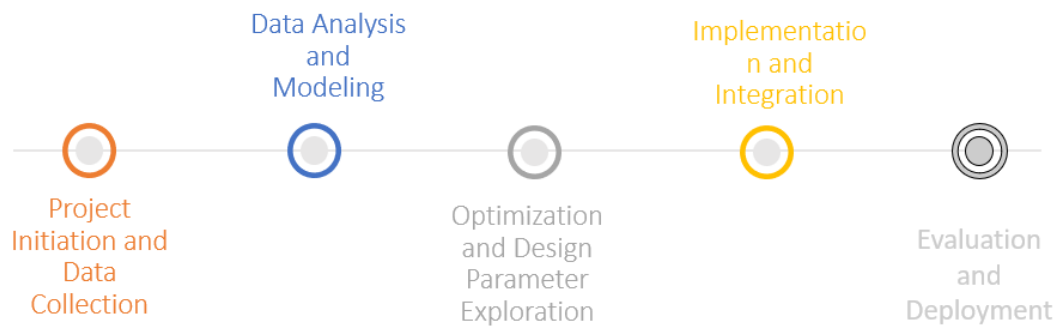
Implementation: Design and implement a database schema adhering to 2/3NF normalization principles for efficient data operations.

6. Git:

Description: Git is a version control system enabling collaborative development, code integrity, and effective project management.

Implementation: Use Git for version control, branching, and collaboration among team members.

4.4 PROJECT MANAGEMENT PLAN



4.5 PROJECT PLAN TIMELINE

Phase 1: Project Initiation and Planning

Week 1-2: Define Objectives and Scope

Define project objectives and scope

Identify key stakeholders

Set project milestones

Week 3-4: Requirements Gathering

Specify web scraping requirements

Define data storage and processing needs

Outline normalization and data modeling requirements

Identify visualization and insights goals

Phase 2: Data Collection and Storage

Week 5-6: Web Scraping Implementation

Develop web scraping scripts

Test and refine web scraping processes

Ensure data quality and integrity

Week 7-8: Azure Data Lake and Azure SQL Database Setup

Provision Azure Data Lake storage

Configure Azure SQL Database

Implement data pipeline using Azure Data Factory

Phase 3: Data Processing and Transformation

Week 9-10: Normalization and Data Modeling

Apply normalization techniques to enhance data structure

Develop a comprehensive data model

Ensure data consistency and integrity

Phase 4: Data Visualization and Analysis

Week 11-12: Data Visualization Tools Integration

Choose appropriate data visualization tools (e.g., Power BI)

Develop visualization dashboards

Test and refine visualizations

Phase 5: Insights Generation

Week 13-14: Derive Insights from Data

Apply analytical techniques to extract meaningful insights

Validate insights against project objectives

Document findings and recommendations

Phase 6: Documentation and Reporting

Week 15-16: Project Documentation

Prepare comprehensive project documentation

Document data engineering processes

Create user guides for ongoing maintenance

4.6 MODULES

- Data Collection (Web Scrapping)
- Integration of raw data in Bronze Layer
- Data Processing to Silver Layer
- Data Aggregation to Gold Layer
- Power App Canvas Application
- Power BI Report

REFERENCES

- [1] Kutay, J. (2022). Data Warehouse vs. Data Lake vs. Data Lakehouse: An Overview of Three Cloud Data Storage Patterns.
- [2] [cit. 12.03.2022] <https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview/>
- [3] Armbrust, M.; Ghodsi, A.; Xin, R.; Zaharia, M. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. 11th Annual Conference on Innovative Data Systems Research (CIDR '21), January 11–15, 2021, Online.
- [4] https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf
- [5] Ravat, F.; Zhao, Y. (2019). Metadata management for data lakes. European Conference on Advances in Databases and Information Systems (ADBIS 2019), vol. 1064, pp. 37–44. Springer, Cham. https://doi.org/10.1007/978-3-030-30278-8_5
- [6] Azeroual, O.; Saake, G.; Abuosba, M.; Schöpfel, J. (2019). Solving Problems of Research Information Heterogeneity During Integration
- [7] Gallinucci, E., Golfarelli, M., Rizzi, S., 2018b. Variety-aware olap of document-oriented databases, in: DOLAP, Vienna, Austria.
- [8] Alqarni, A.A., Pardede, E., 2012. Integration of data warehouse and unstructured business documents. 2012 15th International Conference on Network-Based Information Systems , 32–37.
- [9] Kandel, S.; Heer, J.; Plaisant, C.; Kennedy, J.; van Ham, F.; Riche, N.H.; Weaver, C.; Lee, B.; Brodbeck, D.; Buono, P. (2011).
- [10] Researchdirections in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271
- [11] <https://doi.org/10.1177/1473871611415994>
- [12] Low, W. L.; Lee, M. L.; Ling, T. W. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Information*
- [13] Wang, R.Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2): 58–65. <https://doi.org/10.1145/269012.269022>