

A project report on

DATA SCIENCE WITH AZURE CLOUD SERVICES – INTERNSHIP AT DELOITTE

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology – Electronics and Communication Engineering

by

ANDREW JOHN J – 18BEC1278



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ELECTRONICS ENGINEERING (SENSE)

June 2022

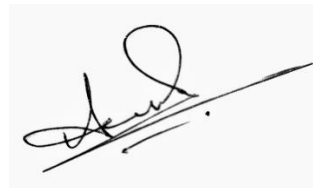
DECLARATION

I hereby declare that the thesis entitled “Data Science with Azure Cloud Services – Internship at Deloitte” submitted by me, for the award of the degree of Bachelor of Technology – Electronics and Communication Engineering, VIT is a record of bona fide work carried out by me under the supervision of Deloitte – USI: Deloitte Support Services India Private Limited.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: 13/06/2022

A handwritten signature in black ink, featuring a large, stylized 'S' or 'D' shape followed by a horizontal line and a small dot.

Signature of the Candidate



Deloitte Support Services India Private Limited
Floor No 15, Deloitte Tower - 1, Survey No 41,
Gachibowli Village, Ranga Reddy District, Hyderabad -
500032, Telangana, India

Tel: +91 040 67621000
www.deloitte.com

May 13, 2022

Mr. Andrew John J

E405, Radiance Mandarin, No.1, Thoraipakkam-Pallavaram 200ft Radial Road, Thoraipakkam,
Chennai, Tamil Nadu, 600097
India

Subject: Offer of Employment

Dear Andrew John J:

On behalf of **Deloitte Support Services India Private Limited** (the "Employer" or "Company"), I am pleased to confirm our offer of employment to you as **Analyst - Deloitte Application Studios** based in **Hyderabad**.

In accordance with the level mapping, your position with the Employer as **Analyst - Deloitte Application Studios** is closely aligned with the position of **Career Level 3** of the Employer. This title alignment is provided for informational purposes only and does not create any additional benefit, entitlement or obligation with regard to your employment with the Employer.

We extend this offer, and the opportunity it represents, with great confidence in your abilities. You have made a very favorable impression with everyone you met and we are excited with the prospect of you joining our organization on **June 20, 2022**.

Your immediate manager will communicate details of your role and work responsibilities in the initial weeks of your joining the Employer. During your employment, the Company may require you to work on any project that you are assigned to, on any technical platforms/skills and nature of the project, in differentiated work timing, at designated work space and location as may be decided by the company.

As part of your annual compensation, you will receive a Total Salary of **Rs. [REDACTED]** and, will be eligible for a performance linked variable bonus. At your level, the variable bonus opportunity could range from **0-10%** of your Total Salary. The actual paid amount could vary depending upon the business and individual performance each fiscal year and, in some situations, could exceed the payout range indicated. Any amounts paid will be subject to statutory and other deductions as per Employer policies and practices. The details of your compensation breakdown are provided in the attached Annexure A.

As an incentive to join the Company, you are eligible to receive a joining bonus of **Rs. [REDACTED]** subject to your reporting for full-time employment on **June 20, 2022**. This amount will attract applicable taxes and will be processed as part of your first month's payroll. You will have an obligation to repay the entire amount of your joining bonus if you resign your position or are terminated for cause by the Company within **12 months** of your start date.

You may also receive additional benefits, including and not limited to, in cash and/or in kind and/or as reimbursement, which could be referred as rewards, awards and gifts, which are generally accorded to the employees of the Employer, subject to the applicable taxes, policies and practices of the Employer.

Your employment with us will be governed by the Terms and Conditions as detailed in **Annexure B**, as well as any and all rules, regulations, guidelines, policies and practices of the Employer, which may be amended from time to time. Deloitte LLP and its U.S.-based subsidiaries (the "Deloitte U.S. Firms") requires their employees to make the necessary representations regarding independence and other matters. Because the Employer is an Indian subsidiary of Deloitte LLP, we must also comply with these independence requirements. Accordingly, this offer is conditional upon you agreeing to make such representations under the Employer's Independence Representations requirements, as further explained in **Annexure B**.

Your compensation details are confidential, and you may discuss it only with the undersigned in case of any clarification. It is our hope that your acceptance of our offer will be just the beginning of a mutually beneficial relationship with our organization. We would like you to join the Employer on **June 20, 2022**, or an alternative mutually agreed upon date.

This offer letter, together with the Annexures described herein, and the Non-Disclosure, Non Solicit and Intellectual Property Rights Assignment Agreement, the Information Security Policy (which you are required to sign upon joining), constitute the entire agreement between the parties with respect to the subject matter of this offer, and supersedes all other previous or contemporaneous oral or written representations, understandings or agreements relating to the subject matter of this offer between you and the Employer or its affiliates.

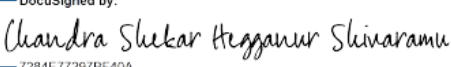
In compliance with applicable laws, Deloitte India (Offices of the US) provides its professionals with home pick-up and drop transport services if their shift timings are between 8:30 pm - 6:00 am in Hyderabad; 9:30 pm - 6:00 am in Mumbai; and 8:00 pm - 6:00 am in Delhi and Bengaluru. Additionally, in Mumbai and Delhi, the firm also provides day-transportation services from central locations to the office and back, at time periods other than those stated above and the associated costs for this conveyance allowance is [REDACTED] deducted on a monthly basis from the payroll, for professionals choosing to opt for the service.

This letter and **Deloitte Support Services India Private Limited** employment application are intended to be final. To accept the offer and the terms of this letter, please sign below in the space provided within three business days.

Andrew John, everyone you have interviewed with joins me in extending to you congratulations and warm regards. We look forward to you joining our team.

Sincerely,

For Deloitte Support Services India Private Limited
Best regards,

DocuSigned by:

7284E77297BF40A...
By: _____
Signature

Authorized Signatory

Acceptance

I, **Andrew John**, hereby accept the terms and conditions of this employment offer.

Please sign and date your Acceptance

DocuSigned by:

80D7AB3FC7D7412...
Signature _____

May 15, 2022
Date _____

ABSTRACT

The world is changing in response to current trends, and Data Science is one such trend in the modern world. For today's youngsters, it is one of the most sought-after employment possibilities. Every business, from large corporations to small startups, requires a Data Scientist to make appropriate use of the massive amounts of data it creates and keeps. Data Science has a wide range of applications in both current and future circumstances. Data science is an interdisciplinary subject that use scientific techniques, procedures, algorithms, and systems to extract information and insights from noisy, structured, and unstructured data, as well as to apply that knowledge across a variety of application areas.

Deloitte Touche Tohmatsu Limited, or Deloitte, is a British multinational professional services firm with operations in over 150 countries and territories worldwide. With headquarters in London, England, Deloitte is one of the Big Four accounting firms and the world's biggest professional services network by revenue and number of professionals. William Welch Deloitte founded the firm in London in 1845, and it extended to the United States in 1890. With over 334,800 experts worldwide, Deloitte provides audit, consulting, financial advice, risk advisory, tax, and legal services.

The name of the Deloitte - USI entity I interned at is Deloitte Support Services India Private Limited. My internship provided me with valuable insights, ideas, and a tremendous amount of corporate and technical experience. My internship at Deloitte lasted exactly 21 weeks, or around 5 months, from January 10th to June 3rd. My internship stipend was ₹25,000. HITEC City in Hyderabad, Telangana, was the place of my internship.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Deloitte – USI: Deloitte Support Services India Private Limited for their constant guidance, continual encouragement, understanding; more than all, they taught me patience in my endeavor. My association with them is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Data Science and Data Engineering.

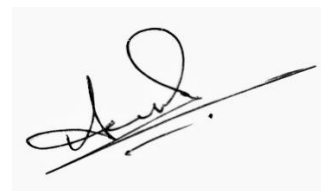
I would like to express my gratitude to Dr. Viswanathan G, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Mr. G.V. Selvam, Dr. Rambabu Kodali, Dr. V.S. Kanchana Bhaaskaran, and Dr. Sivasubramanian A, SENSE, for providing with an environment to work in and for his inspiration during the tenure of the course.

In jubilant mood I express ingeniously my whole-hearted thanks to Dr. Vetrivelan P - Assistant Professor, all teaching staff and members working as limbs of our university for their not-self-centered enthusiasm coupled with timely encouragements showered on me with zeal, which prompted the acquirement of the requisite knowledge to initialize my course study successfully. I would like to thank my parents for their support.

It is indeed a pleasure to thank my friends who persuaded and encouraged me to take up and complete this task. At last, but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly toward the successful completion of this capstone project.

Place: Chennai

Date: 13/06/2022

A handwritten signature in black ink, appearing to read 'Andrew John J', with a long horizontal stroke extending to the right.

Andrew John J

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	DECLARATION	<i>ii</i>
	INTERNSHIP COMPLETION CERTIFICATE	<i>iii</i>
	ABSTRACT	<i>v</i>
	ACKNOWLEDGEMENT	<i>vi</i>
	TABLE OF CONTENTS	<i>vii</i>
	LIST OF FIGURES	<i>ix</i>
	LIST OF TABLES	<i>x</i>
	LIST OF ABBREVIATIONS	<i>xi</i>
	INTRODUCTION	<i>1</i>
1	1.1 ABOUT INTERNSHIP AT DELOITTE	<i>1</i>
	1.2 ABOUT DELOITTE	<i>1</i>
	1.3 ABOUT DATA SCIENCE AND DATA ENGINEERING	<i>3</i>
	1.4 ABOUT TECHNOLOGIES USED, DATA WAREHOUSE, DATA LAKE AND DATA LAKEHOUSE	<i>4</i>
	1.5 OBJECTIVE	<i>5</i>
	BACKGROUND OF THE TECHNOLOGIES USED	<i>6</i>
2	2.1 DATA SCIENCE USED IN VARIOUS APPLICATIONS AND SECTORS OF INDUSTRY	<i>6</i>
	2.2 LITERATURE SURVEY OF APPLICATIONS OF DATA SCIENCE IN RESEARCH	<i>9</i>

	SOFTWARE TECHNOLOGIES USED	14
3	3.1 AZURE DATA FACTORY	14
	3.1.1 WORKING OF ADF	15
	3.1.2 COMPONENTS OF ADF	17
	3.2 AZURE DATABRICKS	20
	3.2.1 DELTA TABLES & DELTA LAKE	21
	3.2.2 DATA LAKEHOUSE VS. DATA WAREHOUSE VS. DATA LAKE	21
	3.2.3 DATABRICKS ARCHITECTURE	23
	3.2.4 ENVIRONMENTS IN DATABRICKS	25
	3.3 DATAKU	27
	3.3.1 DATA, DATASETS & RECIPES	28
	3.4 ROLES & RESPONSIBILITIES OF A DATA SCIENTIST	29
	3.4.1 DIFFERENCE BETWEEN DATA SCIENTIST, DATA ANALYST, AND DATA ENGINEER	30
4	CONCLUSION	31
	4.1 CONCLUSION	31
	4.2 FUTURE SCOPE	31
	REFERENCE	32

LIST OF FIGURES

S. NO	DESCRIPTION OF FIGURE	PAGE NO.
1	Code-Free ETL as a Service	15
2	Data Warehousing	22
3	Data Lake	22
4	Data Lakehouse	23
5	High-Level Databricks Architecture	24

LIST OF TABLES

S. NO	DESCRIPTION OF TABLE	PAGE NO.
1	Difference between DS, DA and DE	30

LIST OF ACRONYMS

FULL FORM	ACRONYM
Azure Data Factory	ADF
Data Warehouse/Warehousing	DW
Enterprise Data Warehouse/Warehousing	EDW
Data Lakehouse/Lakehousing	DL
Data Scientist	DS
Data Analyst	DA
Data Engineer	DE
Systems, Applications & Products	SAP
Extracts, Loads, and Transforms	ELT
Extracts, Transforms, and Loads	ETL
Structured Query Language	SQL
File Transfer Protocol	FTP
Continuous Integration & Continuous Delivery/Deployment	CI/CD
Atomicity, Consistency, Isolation, and Durability	ACID
Business Intelligence	BI
Artificial Intelligence	AI
Machine Learning	ML
Data Science Studio	DSS
Database	DB
US Offices in India	USI

Chapter 1

Introduction

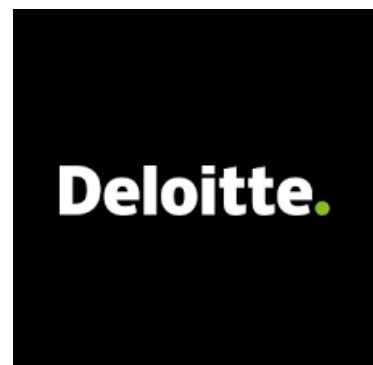
1.1 ABOUT INTERNSHIP AT DELOITTE



The Analyst Internship at Deloitte is the position on which I have spent the most of my time. As an Analyst, I developed and maintained solutions based on technologies such as **SAP**, **Microsoft**, Lotus Notes Dominos, Service Desk / **Service Now**, and Essbase. I supported my team, which uses a simplified system development technique to deliver best-of-breed solutions to clients. I was assigned to the following positions while learning various development tools, testing tools, techniques, and processes: Specialist. Azure Data Factory (ADF), Azure Databricks, and Dataiku are the technologies on which I've been working on extensively. These technologies are more concentrated on Data Warehousing, Data Lake, and Data Lakehouse. My internship at Deloitte lasted from January 10th till June 3rd, precisely 21 weeks or roughly 5 months of time. My internship location was at Hyderabad, Telangana.

1.2 ABOUT DELOITTE

Deloitte Touche Tohmatsu Limited, commonly referred to as Deloitte, is a British multinational professional services company which administers a network with offices in over 150 countries and territories around the world. Deloitte is one of the Big Four accounting organizations and the largest professional services network in the world by revenue and number of professionals, with headquarters in London, England. Deloitte provides audit, consulting, financial advisory, risk advisory, tax, and legal services. Deloitte member firms offer services in the following functions, with country-specific variations on their legal implementation (i.e., all operating within



a single company or through separate legal entities operating as subsidiaries of an umbrella legal entity for the country),

- ✚ Audit provides the organization's traditional accounting and audit services, as well as internal auditing, IT control assurance and Media & Advertising Assurance.
- ✚ Consulting assists clients by providing services in the offering areas of Strategy, Analytics and M&A, Customer and Marketing, Core Business Operations, Human Capital, and Enterprise Technology and Performance. Consulting is Deloitte's largest business.
- ✚ Financial advisory provides corporate finance services to clients, including dispute, personal and commercial bankruptcy, forensics, e-discovery, document review, advisory, mergers & acquisitions, capital projects consulting and valuation services.
- ✚ Risk advisory provides offerings in enterprise risk management, information security and privacy, data quality and integrity, strategic & reputation risk, regulatory risk, project risk and cyber risk, and business continuity management and sustainability.
- ✚ Tax and legal helps clients increase their net asset value, undertake the transfer pricing and international tax activities of multinational companies, minimize their tax liabilities, implement tax computer systems, and provides advisory of tax implications of various business decisions.
- ✚ GovLab is the internal think tank of Deloitte Consulting LLP's Federal Government consulting practice, focused on innovation and government reform.

In India, Deloitte has two entities: Deloitte India and Deloitte US-India (USI), which is a region within the Deloitte US organization. Deloitte India caters to clients within India, while Deloitte USI is an entity of Deloitte US that is geographically located in India and caters to clients of the US member firm.

1.3 ABOUT DATA SCIENCE AND DATA ENGINEERING

Data is the new Oil. This statement shows how every modern IT system is driven by capturing, storing and analyzing data for various needs. Be it about making decision for business, forecasting weather, studying protein structures in biology or designing a marketing campaign. All of these scenarios involve a multidisciplinary approach of using mathematical models, statistics, graphs, databases and of course the business or scientific logic behind the data analysis. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Data engineering is the complex task of making raw data usable to data scientists and groups within an organization. Data engineering encompasses numerous specialties of data science. In addition to making data accessible, data engineers create raw data analyses to provide predictive models and show trends for the short- and long-term. Without data engineering, it would be impossible to make sense of the huge amounts of data that are available to businesses. There are four key phases of the data pipeline that data engineering directly deals with:

- ✚ **Ingestion** - This is the task of gathering data. Depending on the number of data sources, this task can be focused or large-scale.
- ✚ **Processing** - During this phase, ingested data is sorted to achieve a specific set of data to analyze. For large data sets, this is commonly done using a distributed computing platform for scalability.
- ✚ **Storing** - This takes the results of the processing and saves the data for fast and easy retrieval. The effectiveness of this phase relies on a sound database management system - which can be on premise or in the cloud
- ✚ **Access** - Once in place, the data is available to users with access.

1.4 ABOUT TECHNOLOGIES USED, DATA WAREHOUSE, DATA LAKE AND DATA LAKEHOUSE

The technologies that were used in my internship were Azure Data Factory, Azure Databricks and Dataiku. Apart from the 3 listed



technologies, we also work on ServiceNow, SAP and Microsoft Office 365. The technologies used had the essence of Data Science and Data Engineering in them, but these were the foundational layers to Data Warehousing, which was the main work our team did. Data warehouses and data lakes have been the most widely used storage architectures for big data. A **Data Warehouse** (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis and is considered a core component of business intelligence. Data Warehouses are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place that are used for creating analytical reports for workers throughout the enterprise. A **Data Lake** is a system or repository of data stored in its natural/raw format, usually object blobs or files. A data lake is usually a single store of data including raw copies of source system data, sensor data, social data etc., and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, and video). A data lake can be established "on premises" or "in the cloud". A **Data Lakehouse** is a new data storage architecture that combines the flexibility of data lakes and the data management of data warehouses.

Azure Data Factory is Azure's cloud ETL service for scale-out serverless data integration and data transformation. It offers a code-free UI for intuitive authoring and single-pane-of-glass monitoring and management. We can also lift and shift existing SQL Server Integration Services packages to Azure and run them with full compatibility in ADF. SSIS Integration Runtime offers a fully managed service, so we don't have to worry about infrastructure management. **Databricks** is a cloud-

based collaborative data science, data engineering, and data analytics platform that combines the best of data warehouses and data lakes into a lakehouse architecture. **Azure Databricks** is a data analytics platform optimized for the Microsoft Azure cloud services platform. Azure Databricks offers three environments for developing data intensive applications: Databricks SQL, Databricks Data Science & Engineering, and Databricks Machine Learning.

1.5 OBJECTIVE




The objective is to learn and apply Data Science & Data Engineering, and to use the core concepts to build Data Warehouse and develop Data Lake, and thereafter integrate them to build Data Lakehouse using Azure Data Factory, Azure Databricks and Dataiku.

Chapter 2


Background of the Technologies Used


2.1 DATA SCIENCE USED IN VARIOUS APPLICATIONS AND SECTORS OF INDUSTRY


Data Science is the deep study of a large quantity of data, which involves extracting some meaningful from the raw, structured, and unstructured data. The extracting out meaningful data from large amounts use processing of data and this processing can be done using statistical techniques and algorithm, scientific techniques, different technologies, etc. It uses various tools and techniques to extract meaningful data from raw data. Data Science is also known as the Future of Artificial Intelligence. Some of the major applications and some of the uses in some sectors of industries are listed below:

-  **In Search Engines:** The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.
-  **Autocomplete:** AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line. In Google Mail, when we are writing formal mail to someone so at that time data science concept of Autocomplete feature is used where he/she is an efficient choice to auto-complete the whole line. Also, in Search Engines in social media, in various apps, AutoComplete feature is widely used.
-  **Fraud and Risk Detection:** Finance made was an entrant in the data applications. Finance and data science go hand in hand as, like data science, Finance is also about data. Earlier companies had loads of paperwork to initiate sanctioning loans, maintaining them, incurring losses, and being in debt. So data science practices were thought of as the solution. To analyze risk probabilities, they learned to separate the data by customer profiling, past expenditures, and

other necessary variables. It also helps to push their banking products based on customer's purchasing power. Another application could be that customer portfolio management analyzes trends in data through business intelligence tools for data science. Data science also introduces algorithmic training; financial institutions can make data-driven decisions through rigorous data analysis. Therefore, making the customer experiences better for the users as through extensive analysis of client experience and modification of preferences, financial institutions can create a personalized relationship with their customers.

 **Image Recognition and Speech Recognition:** Data science algorithms rule the speech and image recognition fields. We can come across the great work of these algorithms in our everyday lives. Siri, Alexa, and Google Assistant has its speech recognition algorithm is working behind the system, trying to understand and evaluating our words and returning with productive outcomes of our use. Image recognition can also be all over our social media sites like Facebook, Instagram, and Twitter. These apps offer to recognize the person on our list and offer to tag them when we upload a picture with them on our profile.

 **Medicine and Drug Development:** The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors. The algorithms based on data science will forecast how this will react to the human body without lab experiments.

 **Airline Route Planning:** Companies like Southwest Airlines, Alaska Airlines have started to incur data science in their processing of flights to bring in the new change in their working style. Earlier airline companies suffered a massive loss with rising fuel prices as it became difficult for them to maintain occupancy ratio and operating profits. With data science, airline companies can think of strategic improvements like predicting flight delays, deciding on the aircraft to purchase, planning routes and layovers, and marketing strategies like a customer

loyalty program.


✚ **Transportation:** The most significant advancement or the evolution that data science has given us in transportation is introducing self-driving cars. Data science has created a stronghold in transport through extensive analysis of fuel consumption patterns, driver behavior, and vehicle monitoring. It is making its name by providing safer driving environments for drivers, optimizing vehicle performance, adding autonomy to the driver, and much more. Through reinforcement learning and introduction to autonomy, vehicle manufacturers can create intelligent automobiles and better logistical routes. Popular cab services like Uber use data science to use various variables such as customer profiles, location, economic indicators, and logistics vendors to optimize price and delivery routes and proper allocation of resources.

✚ **Gaming:** There are also machine and data science algorithms that upgrade themselves to a new level as the gamer moves up the higher level in the game. The algorithm is designed and developed to analyze the previous performance of the gamer and shape up the game accordingly. Top-notch gaming studios like Zynga, EA Sports have upgraded to a new experience altogether with the help of such algorithms.

✚ **E-commerce:** Data science algorithms and machine learning concepts like NLP and recommendation systems are hugely benefitting the e-commerce market. The e-commerce platforms can study customer purchases and feedback using such techniques to get powerful insights for their business development. They make use of NLP to analyze texts and online surveys. It is used in collaborative and content-based filtering to analyze the data and provide better services to its customers. Other ways that data science has impacted the data science industry include identifying customer base, forecasting goods and services, identifying the style of popular products, optimizing pricing structures, and more.

✚ **Banking:** Data science has enabled banks across the globe to be more secure and manage their resources efficiently. It also enables them to make smarter and more strategic decisions and be saved from fraud. It also helps manage customer

data, risk analysis and modeling, predictive analysis, and much more. Data science allows bankers to assess the customer lifetime value allowing them to monitor and thus derive several predictions and analyze investment patterns of customers for their business. Machine learning algorithms improve analytics strategy in real-time.

 **Manufacturing:** Data scientists are the new factory labor in the manufacturing industry and thus acquired a crucial position in manufacturing and retail. It has shrunk away redundant jobs by introducing powerful machinery using machine learning techniques such as reinforcement learning. Moreover, integrating with technologies like the Internet of Things (IoT) enables the industries to predict potential problems, monitor systems, and analyze the continuous stream of data.

Let us see how does data science impacts manufacturing:

- Optimize energy costs and productive hours.
- Improving decisions and improving the quality of the products based on customer reviews.
- Build an autonomous system using historical and real-time data to boost up the manufacturing line.

2.2 LITERATURE SURVEY OF APPLICATIONS OF DATA SCIENCE IN RESEARCH

ADF, Databricks and Dataiku is used in several industries and all 3 technologies are mainly a large user of core concepts of Data Science, but when it comes it to research, there are still groundbreaking results accomplished to automate more industry processes. Sofie *et al.* [1] have told in their research paper that digital transformation and implementation of big data platforms are inevitable in any industry. Big data constitutes an important area of research, however, implementation of platforms like Microsoft Azure have yet to be explored. Through a narrative case study, they aim to explore the implementation of such big data platforms in the power industry. Their case is based in a Norwegian power company who are early movers in implementing Microsoft's Azure platform across multiple units in the organization. With the support of top management and eager business units one would expect this process to be fairly straight forward. Their findings show that the maturity of the technology, in addition to

challenges of being an early mover, create an unexpected path to success.

Milind *et al.* [2], theorized a working approach, i.e., Data Migration has become important aspect nowadays when it comes to data movement from on premise databases to cloud storage or cloud databases. In this paper we present working case study using cloud based ETL tool known as Azure Data Factory used for Data Migration from on premise Oracle database to cloud-based SQL Managed Instance database for an organization. This paper evaluates the implementation of data migration process in general and specific to the tool and technologies involved in data migration process using Azure Data Factory for an organization. When an organization needs to move their application to cloud, the essential of data migration needs to be discussed, proper architecture is required to further break down each task to migrate the data. The proof of concept should be established to see if data is not getting truncated/altered in the process of migration and existing logic on the on-premises database works well after moving data to the cloud. In this paper we also discuss about encryption process involved while migrating data as this is an important aspect in data migration to migrate data with existing algorithms used in on premise database and its implementation while data movement takes place using Azure Data Factory. In Oracle there are encryption algorithms being used to store sensitive user data, we have to analyze existing encryption/decryption process and implement an architecture with the help of data migration tool so that data remains intact after movement. Developer has to develop testing strategies to compare on premise data versus the data moved to the cloud storage. Azure Data Factory is powerful cloud-ETL tool to move your hundreds of table data at a time to new cloud database with maximum data transfer throughput. This data migration process requires thorough evaluation of multiple factors e.g., actual table size in migration, throughput, Virtual Machine used for data transfer, network bandwidth etc.

Furthermore, on Databricks from a SQL and R standpoint, there has been numerous numbers of papers, Michael *et al.* [3] have said Spark SQL is a new module in Apache Spark that integrates relational processing with Spark's functional programming API. Built on our experience with Shark, Spark SQL lets Spark programmers leverage the benefits of relational processing (e.g. declarative queries and optimized storage), and lets SQL users call complex analytics libraries in Spark (e.g. machine learning).

Compared to previous systems, Spark SQL makes two main additions. First, it offers much tighter integration between relational and procedural processing, through a declarative Data Frame API that integrates with procedural Spark code. Second, it includes a highly extensible optimizer, Catalyst, built using features of the Scala programming language, which makes it easy to add composable rules, control code generation, and define extension points. Using Catalyst, we have built a variety of features (e.g., schema inference for JSON, machine learning types, and query federation to external databases) tailored for the complex needs of modern data analysis. We see Spark SQL as an evolution of both SQL-on-Spark and of Spark itself, offering richer APIs and optimizations while keeping the benefits of the Spark programming model. Shivaram *et al.* [4] have discussed that R is a popular statistical programming language with a number of extensions that support data processing and machine learning tasks. However, interactive data analysis in R is usually limited as the R runtime is single threaded and can only process data sets that fit in a single machine's memory. We present SparkR, an R package that provides a frontend to Apache Spark and uses Spark's distributed computation engine to enable large scale data analysis from the R shell. We describe the main design goals of SparkR, discuss how the high-level Data Frame API enables scalable computation and present some of the key details of our implementation.

As we know Databricks major programming languages are Scala, SQL, R and Python, there has been some good amount of research in python standpoint, where Xiangrui *et al.* [5] presented MLlib, Spark's open-source distributed machine learning library. MLlib provides efficient functionality for a wide range of learning settings and includes several underlying statistical, optimization, and linear algebra primitives. Shipped with Spark, MLlib supports several languages and provides a high-level API that leverages Spark's rich ecosystem to simplify the development of end-to-end machine learning pipelines. MLlib has experienced a rapid growth due to its vibrant open-source community of over 140 contributors and includes extensive documentation to support further growth and to let users quickly get up to speed. Moving to Big Data perspective, Matei *et al.* [6] have briefed that scalable data processing will be essential for the next generation of computer applications but typically involves a complex sequence of processing steps with different computing systems. To simplify this task, the Spark project introduced a unified programming model and engine for big data applications.

Their experience shows such a model can efficiently support today's workloads and brings substantial benefits to users. They hope Apache Spark highlights the importance of composability in programming libraries for big data and encourages development of more easily interoperable libraries.

Moving to a cluster standpoint that are used to run jobs in Databricks, Matei *et al.* [7] discusses about MapReduce and its variants that have been highly successful in implementing large-scale data intensive applications on clusters of unreliable machines. However, most of these systems are built around an acyclic data flow programming model that is not suitable for other popular applications. In this paper, they focus on one such class of applications: those that reuse a working set of data across multiple parallel operations. This includes many iterative machine learning algorithms, as well as interactive data analysis environments. They propose a new framework called Spark that supports these applications while maintaining the scalability and fault-tolerance properties of MapReduce. To achieve these goals, Spark introduces a data abstraction called resilient distributed datasets (RDDs). An RDD is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost. Spark can outperform Hadoop by 10x in iterative machine learning jobs and can be used to interactively query a 39 GB dataset with sub-second response time. In [8], Michael *et al.*, discuss about the usability and performance of scaling Spark in the real world where, on the usability side, they are augmenting Spark with a large set of standard libraries containing scalable versions of common data analysis algorithms. For example, Spark's machine learning library, MLlib, grew by a factor of 4 in the past year. They have also designed a pluggable data source API that makes it easy to access external data sources in a uniform way using DataFrames or SQL. Together, these APIs form one of the largest integrated standard libraries for "big data," and will undoubtedly lead to interesting design decisions to enable efficient composition of workflows. On the performance side, under a new project codenamed Tungsten, they are implementing memory management outside the JVM and runtime code generation to bring the performance of DataFrames and SQL to the limit of the underlying hardware [11]. These optimizations will transparently speed up current user code and many of Spark's libraries. They have also increasingly seen Spark used in research projects, including online aggregation [12], graph processing [13], genomic data processing [14], and large-scale neuroscience [15]. They hope that Spark's relatively small code size and

wide array of built-in functions make it amenable to both systems and application-oriented projects.

On the cloud computing integration with data science standpoint, Michael *et al.* [9] presume and predict cloud computing will grow, so developers should take it into account. Regardless of whether a cloud provider sells services at a low level of abstraction like EC2 or a higher level like AppEngine, they believe computing, storage, and networking must all focus on horizontal scalability of virtualized resources rather than on single node performance. Moreover: 1. Applications software needs to both scale down rapidly as well as scale up, which is a new requirement. Such software also needs a pay-for-use licensing model to match needs of cloud computing. 2. Infrastructure software must be aware that it is no longer running on bare metal but on VMs. Moreover, metering and billing need to be built in from the start. 3. Hardware systems should be designed at the scale of a container (at least a dozen racks), which will be the minimum purchase size. Cost of operation will match performance and cost of purchase in importance, rewarding energy proportionality by putting idle portions of the memory, disk, and network into low-power mode. Processors should work well with VMs, and flash memory should be added to the memory hierarchy, and LAN switches and WAN routers must improve in bandwidth and cost.

On Dataiku, which is an up-and-coming AI/ML-Data Science tool, where Catalina Herrera, A Dataiku Senior Engineer, presented a webinar [10], “Breaking Silos: the power of collaboration and abstraction”. This webinar introduced examples of a community working together to deliver a common end result using (and reusing data) to maximize community outcomes. It also presented how to leverage Dataiku’s capabilities to a wide spectrum of applications, including Data4Good, wind turbines with public data, Co2 emissions, among others, to help drive understanding of how to deliver and consume data and insights from many diverse data sources, including observations, model predictions, and experts’ knowledge (i.e. evidence). The presentation showed how once we break down silos, it’s important to enhance data products through collaboration, and leverage Machine Learning / Artificial Intelligence to deliver applied data science as transparent consumables.

Chapter 3

Software Technologies Used

3.1 AZURE DATA FACTORY

In the world of big data, raw, unorganized data is often stored in relational, non-relational, and other storage systems.



However, on its own, raw data doesn't have the proper context or meaning to provide meaningful insights to analysts, data scientists, or business decision makers. Big data requires a service that can orchestrate and operationalize processes to refine these enormous stores of raw data into actionable business insights. Azure Data Factory is a managed cloud service that's built for these complex hybrid extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects.

Considering a usage scenario, for example, imagine a gaming company that collects petabytes of game logs that are produced by games in the cloud. The company wants to analyze these logs to gain insights into customer preferences, demographics, and usage behavior. It also wants to identify up-sell and cross-sell opportunities, develop compelling new features, drive business growth, and provide a better experience to its customers. To analyze these logs, the company needs to use reference data such as customer information, game information, and marketing campaign information that is in an on-premises data store. The company wants to utilize this data from the on-premises data store, combining it with additional log data that it has in a cloud data store. To extract insights, it hopes to process the joined data by using a Spark cluster in the cloud (Azure HDInsight) and publish the transformed data into a cloud data warehouse such as Azure Synapse Analytics to easily build a report on top of it. They want to automate this workflow and monitor and manage it on a daily schedule. They also want to execute it when files land in a blob store container.

Azure Data Factory is the platform that solves such data scenarios. It is the cloud-based ETL and data integration service that allows us to create data-driven workflows for orchestrating



data movement and transforming data at scale. Using Azure Data Factory, we can create and schedule data-driven workflows (called pipelines) that can ingest data from

disparate data stores. We can build complex ETL processes that transform data visually with data flows or by using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure SQL Database. Additionally, we can publish your transformed data to data stores such as Azure Synapse Analytics for business intelligence (BI) applications to consume. Ultimately, through Azure Data Factory, raw data can be organized into meaningful data stores and data lakes for better business decisions.

3.1.1 WORKING OF ADF

Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers.

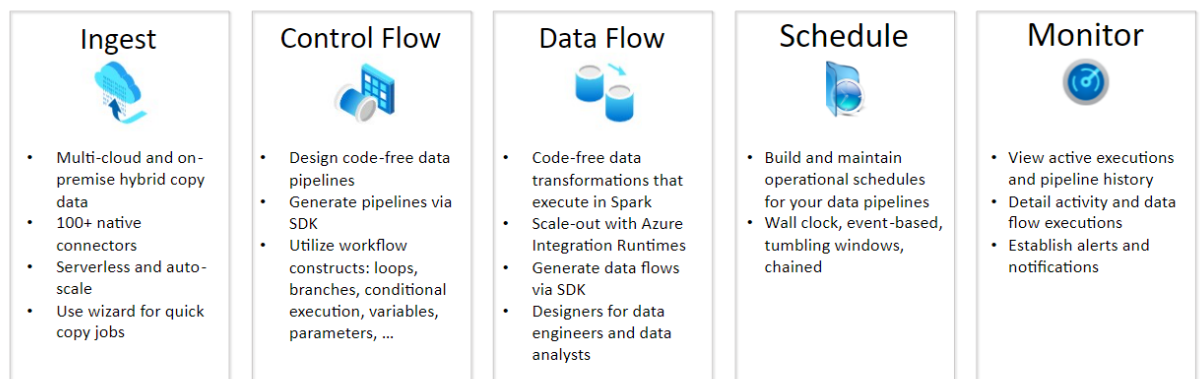


Figure 1: Code-Free ETL as a Service

Connect and Collect:

- Enterprises have data of various types that are located in disparate sources on-premises, in the cloud, structured, unstructured, and semi-structured, all arriving at different intervals and speeds.
- The first step in building an information production system is to connect to all the required sources of data and processing, such as software-as-a-service (SaaS) services, databases, file shares, and FTP web services. The next step is to move the data as needed to a centralized location for subsequent processing.
- Without Data Factory, enterprises must build custom data movement components or write custom services to integrate these data sources and processing. It's expensive and hard to integrate and maintain such systems. In addition, they often lack the enterprise-grade monitoring,

alerting, and the controls that a fully managed service can offer.

- With Data Factory, you can use the Copy Activity in a data pipeline to move data from both on-premises and cloud source data stores to a centralization data store in the cloud for further analysis. For example, you can collect data in Azure Data Lake Storage and transform the data later by using an Azure Data Lake Analytics compute service. You can also collect data in Azure Blob storage and transform it later by using an Azure HDInsight Hadoop cluster.

Transform and Enrich:

- After data is present in a centralized data store in the cloud, process or transform the collected data by using ADF mapping data flows. Data flows enable data engineers to build and maintain data transformation graphs that execute on Spark without needing to understand Spark clusters or Spark programming.
- If you prefer to code transformations by hand, ADF supports external activities for executing your transformations on compute services such as HDInsight Hadoop, Spark, Data Lake Analytics, and Machine Learning.

CI/CD and Publish:







- Data Factory offers full support for CI/CD of your data pipelines using Azure DevOps and GitHub. This allows you to incrementally develop and deliver your ETL processes before publishing the finished product. After the raw data has been refined into a business-ready consumable form, load the data into Azure Data Warehouse, Azure SQL Database, Azure Cosmos DB, or whichever analytics engine your business users can point to from their business intelligence tools.

Monitor:

- After you have successfully built and deployed your data integration pipeline, providing business value from refined data, monitor the scheduled activities and pipelines for success and failure rates. Azure Data Factory has built-in support for pipeline monitoring via Azure Monitor, API, PowerShell, Azure Monitor logs, and health panels on the Azure portal

3.1.2 COMPONENTS OF ADF

An Azure subscription might have one or more Azure Data Factory instances (or data factories). Azure Data Factory is composed of below key components.

-  Pipelines
-  Activities
-  Datasets
-  Linked services
-  Data Flows
-  Integration Runtimes

These components work together to provide the platform on which you can compose data-driven workflows with steps to move and transform data.

Pipeline:

- A data factory might have one or more pipelines. A pipeline is a logical grouping of activities that performs a unit of work. Together, the activities in a pipeline perform a task. For example, a pipeline can contain a group of activities that ingests data from an Azure blob, and then runs a Hive query on an HDInsight cluster to partition the data.
- The benefit of this is that the pipeline allows you to manage the activities as a set instead of managing each one individually. The activities in a pipeline can be chained together to operate sequentially, or they can operate independently in parallel.

Mapping data flows:

- Create and manage graphs of data transformation logic that you can use to transform any-sized data. You can build-up a reusable library of data transformation routines and execute those processes in a scaled-out manner from your ADF pipelines. Data Factory will execute your logic on a Spark cluster that spins-up and spins-down when you need it. You won't ever have to manage or maintain clusters.

Activity:

- Activities represent a processing step in a pipeline. For example, you might use a copy activity to copy data from one data store to another data store. Similarly, you might use a Hive activity, which runs a Hive query on an Azure HDInsight cluster, to transform or analyze your data. Data Factory supports three types of activities: data movement activities, data transformation activities, and control activities.

Datasets:

- Datasets represent data structures within the data stores, which simply point to or reference the data you want to use in your activities as inputs or outputs.

Linked services:

- Linked services are much like connection strings, which define the connection information that's needed for Data Factory to connect to external resources. Think of it this way: a linked service defines the connection to the data source, and a dataset represents the structure of the data. For example, an Azure Storage-linked service specifies a connection string to connect to the Azure Storage account. Additionally, an Azure blob dataset specifies the blob container and the folder that contains the data.
- Linked services are used for two purposes in Data Factory:
 - 1) To represent a data store that includes, but isn't limited to, a SQL Server database, Oracle database, file share, or Azure blob storage account. For a list of supported data stores, see the copy activity article.
 - 2) To represent a compute resource that can host the execution of an activity. For example, the HDInsightHive activity runs on an HDInsight Hadoop cluster. For a list of transformation activities and supported compute environments, see the transform data article.

Integration Runtime:

- In Data Factory, an activity defines the action to be performed. A linked service defines a target data store or a compute service. An integration runtime provides the bridge between the activity and linked Services. It's referenced by the linked service or activity and provides the compute environment where the activity either runs on or gets dispatched from. This way, the activity can be performed in the region closest possible to the target data store or compute service in the most performant way while meeting security and compliance needs.

Triggers:

- Triggers represent the unit of processing that determines when a pipeline execution needs to be kicked off. There are different types of triggers for different types of events.

Pipeline runs:

- A pipeline run is an instance of the pipeline execution. Pipeline runs are typically instantiated by passing the arguments to the parameters that are defined in pipelines. The arguments can be passed manually or within the trigger definition.

Parameters:

- Parameters are key-value pairs of read-only configuration. Parameters are defined in the pipeline. The arguments for the defined parameters are passed during execution from the run context that was created by a trigger or a pipeline that was executed manually. Activities within the pipeline consume the parameter values.
- A dataset is a strongly typed parameter and a reusable/referenceable entity. An activity can reference datasets and can consume the properties that are defined in the dataset definition.
- A linked service is also a strongly typed parameter that contains the connection information to either a data store or a compute environment.

Control flow:

- Control flow is an orchestration of pipeline activities that includes chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger. It also includes custom-state passing and looping containers, that is, For-each iterator.

Variables:

- Variables can be used inside of pipelines to store temporary values and can also be used in conjunction with parameters to enable passing values between pipelines, data flows, and other activities.

3.2 AZURE DATABRICKS

Databricks is a cloud-based collaborative data science, data engineering, and data analytics platform that combines the best of data warehouses and data lakes into a lakehouse architecture. The Databricks

Lakehouse combines the ACID transactions and data governance of data warehouses with the flexibility and cost-efficiency of data lakes to enable business intelligence (BI) and machine learning (ML) on all data. The Databricks Lakehouse keeps your data in your massively scalable cloud object storage in open-source data standards, allowing you to use your data however and wherever you want. The primary components of the Databricks Lakehouse are:



databricks

Delta tables:

- ACID transactions
- Data versioning
- ETL
- Indexing



Unity Catalog:

- Data governance
- Data sharing
- Data auditing

By storing data with Delta Lake, you enable downstream data scientists, analysts, and machine learning engineers to leverage the same production data supporting your core ETL workloads as soon as data is processed. Unity Catalog ensures that you have complete control over who gains access to which data and provides a centralized mechanism for managing all data governance and access controls without needing to replicate your data.

3.2.1 DELTA TABLES & DELTA LAKE

Tables created on Databricks use the Delta Lake protocol by default. When you create a new Delta table:

-  Metadata used to reference the table is added to the metastore in the declared schema or database.
-  Data and table metadata are saved to a directory in cloud object storage.

Delta Lake is a key component of the Databricks lakehouse architecture. The Delta table format is a widely used standard for enterprise data lakes at massive scale. Built on the foundation of another open-source format—Parquet—Delta Lake adds advanced features and capabilities that enable additional robustness, speed, versioning, and data-warehouse-like ACID compliance. This is on top of the existing cost benefits of using existing cheap blob storage services. Databricks has built-in support for Delta Lake, and the latest Databricks Runtimes include performance enhancements for even more speed and performance.

3.2.2 DATA LAKEHOUSE VS. DATA WAREHOUSE VS. DATA LAKE

Data warehouses have powered business intelligence (BI) decisions for about 30 years, having evolved as set of design guidelines for systems controlling the flow of data. Data warehouses optimize queries for BI reports but can take minutes or even

hours to generate results. Designed for data that is unlikely to change with high frequency, data warehouses seek to prevent conflicts between concurrently running queries. Many data warehouses rely on proprietary formats, which often limit support for machine learning.

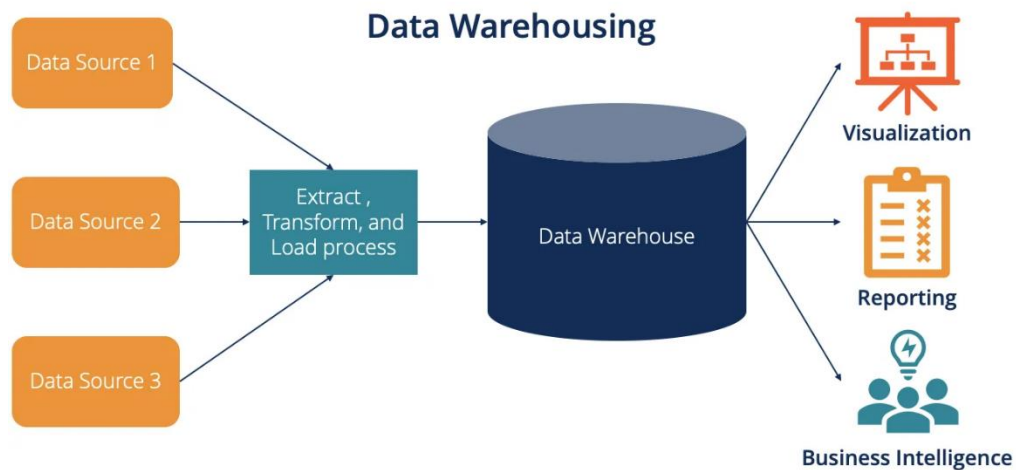


Figure 2: Data Warehousing

Powered by technological advances in data storage and driven by exponential increases in the types and volume of data, data lakes have come into widespread use over the last decade. Data lakes store and process data cheaply and efficiently. Data lakes are often defined in opposition to data warehouses: A data warehouse delivers clean, structured data for BI analytics, while a data lake permanently and cheaply stores data of any nature in any format. Many organizations use data lakes for data science and machine learning, but not for BI reporting due to its unvalidated nature.

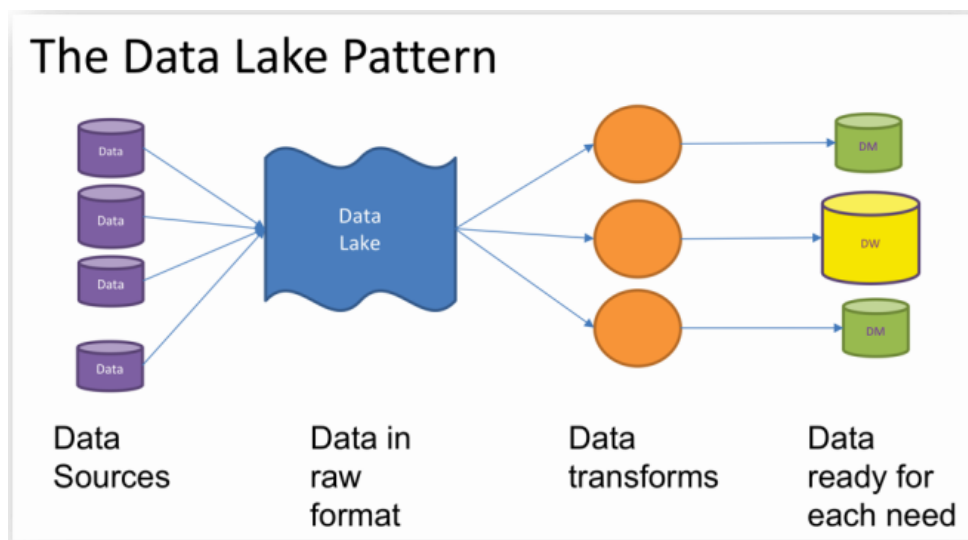


Figure 3: Data Lake

The data lakehouse replaces the current dependency on data lakes and data warehouses for modern data companies that desire:

- 🌈 Open, direct access to data stored in standard data formats.
- 🌈 Indexing protocols optimized for machine learning and data science.
- 🌈 Low query latency and high reliability for BI and advanced analytics.

By combining an optimized metadata layer with validated data stored in standard formats in cloud object storage, the data lakehouse allows data scientists and ML engineers to build models from the same data driving BI reports.

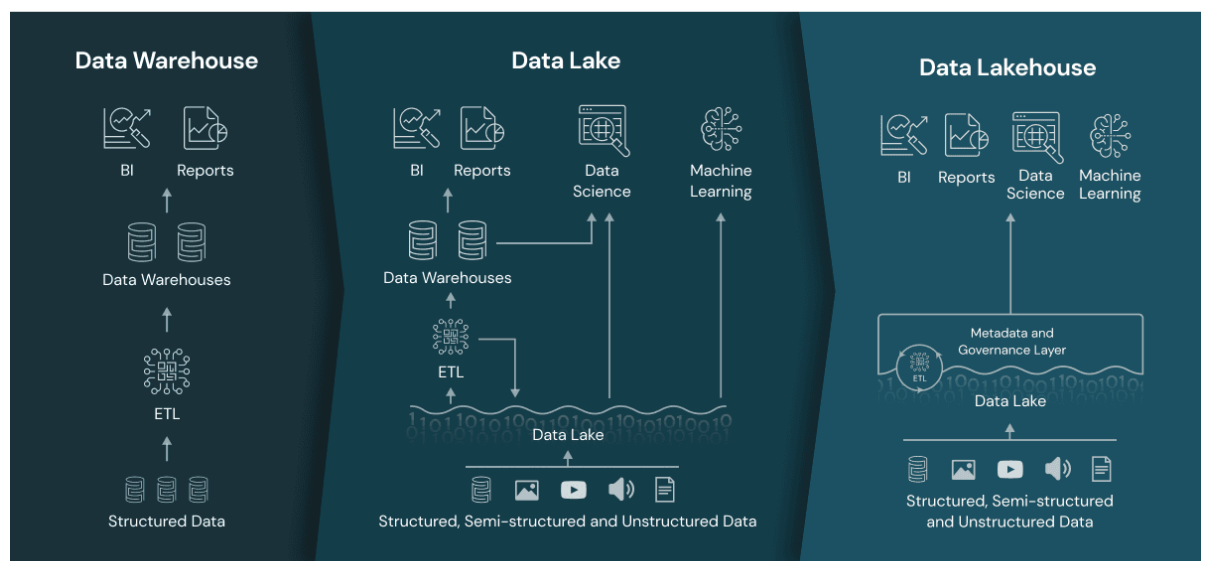


Figure 4: Data Lakehouse

3.2.3 DATABRICKS ARCHITECTURE

Databricks is structured to enable secure cross-functional team collaboration while keeping a significant amount of backend services managed by Databricks so we can stay focused on our data science, data analytics, and data engineering tasks. Databricks operates out of a control plane and a data plane:

- 🌈 The control plane includes the backend services that Databricks manages in its own AWS account. Notebook commands and many other workspace configurations are stored in the control plane and encrypted at rest.

🚦 The data plane is where our data is processed.

- For most Databricks computation, the compute resources are in our AWS account in what is called the Classic data plane. This is the type of data plane Databricks uses for notebooks, jobs, and for Classic Databricks SQL endpoints.
- If we enable Serverless compute for Databricks SQL, the compute resources for Databricks SQL are in a shared Serverless data plane. The compute resources for notebooks, jobs and Classic Databricks SQL endpoints still live in the Classic data plane in the customer account.

The following diagram describes the overall architecture of the Classic data plane.

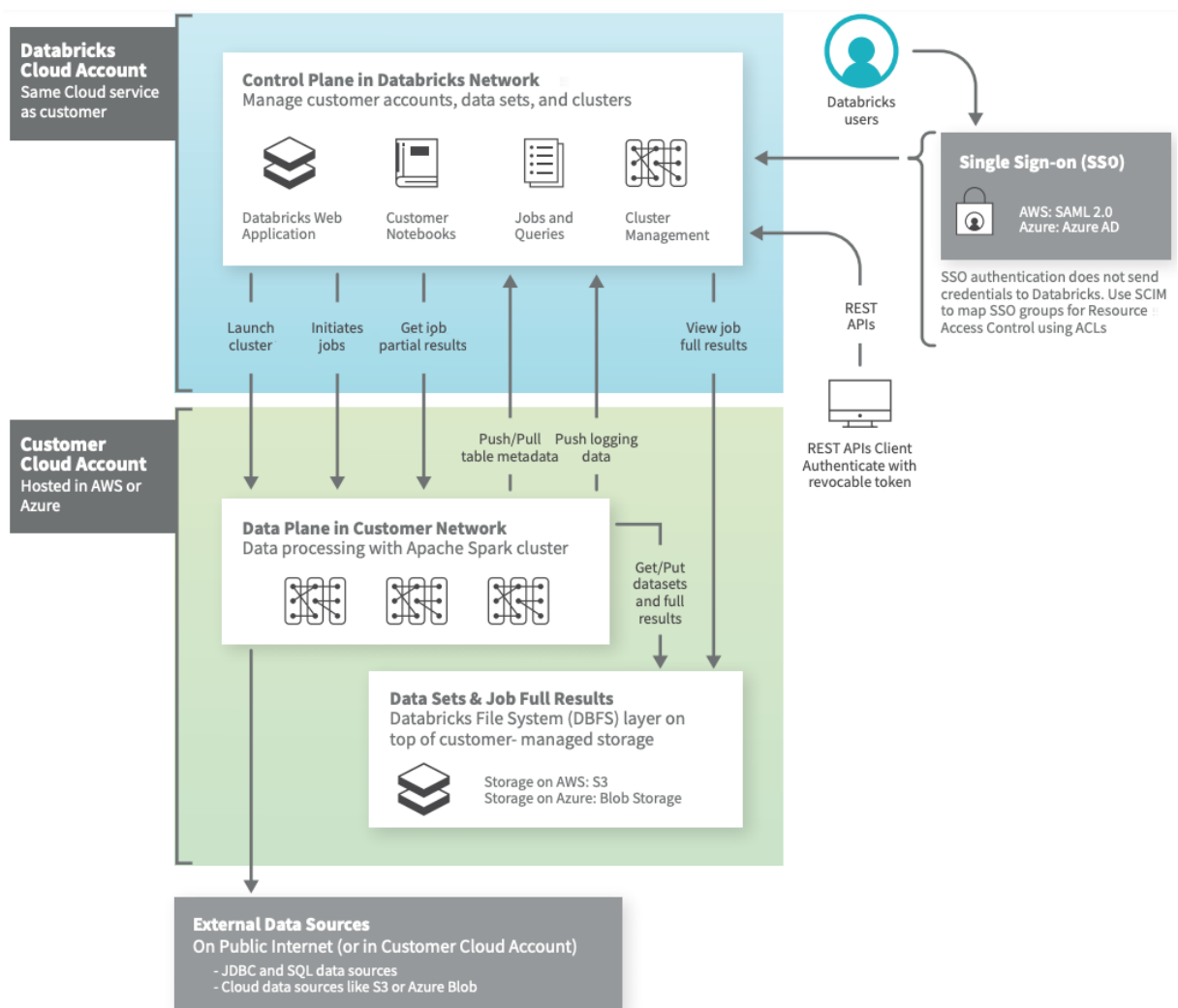
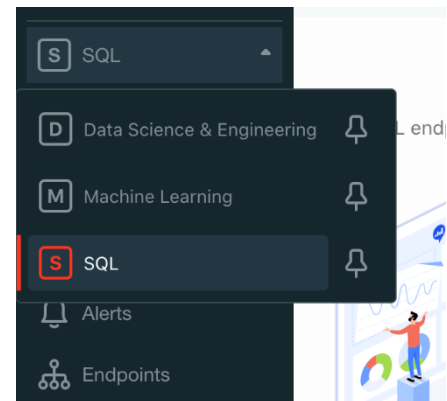


Figure 5: High-Level Databricks Architecture

3.2.4 ENVIRONMENTS IN DATABRICKS

There are 3 major environments Databricks works with, and these are the 3 pillars Databricks operate on, and they are:

- 🔗 Databricks Data Science and Engineering
- 🔗 Databricks Machine Learning
- 🔗 Databricks SQL






Databricks Data Science & Engineering is the classic Databricks environment for collaboration among data scientists, data engineers, and data analysts. A workspace is an environment for accessing all of your Databricks assets. A workspace organizes objects (notebooks, libraries, dashboards, and experiments) into folders and provides access to data objects and computational resources.

Databricks Data Science & Engineering provides an interactive workspace that enables collaboration between data engineers, data scientists, and machine learning engineers. For a big data pipeline, the data (raw or structured) is ingested into Azure through Azure Data Factory in batches, or streamed near real-time using Apache Kafka, Event Hub, or IoT Hub. This data lands in a data lake for long term persisted storage, in Azure Blob Storage or Azure Data Lake Storage. As part of your analytics workflow, use Azure Databricks to read data from multiple data sources and turn it into breakthrough insights using Spark.

- 🔗 **Notebook:** A web-based interface to documents that contain runnable commands, visualizations, and narrative text.
- 🔗 **Dashboard:** An interface that provides organized access to visualizations.
- 🔗 **Library:** A package of code available to the notebook or job running on your cluster. Databricks runtimes include many libraries, and you can add your own.
- 🔗 **Repo:** A folder whose contents are co-versioned together by syncing them to a remote Git repository.
- 🔗 **Experiment:** A collection of MLflow runs for training a machine learning model.

The Databricks Machine Learning environment starts with the features provided in the Data Science & Engineering workspace and adds functionality. Important concepts include:

-  **Experiments:** The main unit of organization for tracking machine learning model development. Experiments organize, display, and control access to individual logged runs of model training code.
-  **Feature Store:** A centralized repository of features. Databricks Feature Store enables feature sharing and discovery across your organization and also ensures that the same feature computation code is used for model training and inference.
-  **Models:** A trained machine learning or deep learning model that has been registered in Model Registry.

Databricks Machine Learning is an integrated end-to-end machine learning environment incorporating managed services for experiment tracking, model training, feature development and management, and feature and model serving.

Databricks SQL is geared toward data analysts who work primarily with SQL queries and BI tools. It provides an intuitive environment for running ad-hoc queries and creating dashboards on data stored in our data lake. Databricks SQL provides an easy-to-use platform for analysts who want to run SQL queries on their data lake, create multiple visualization types to explore query results from different perspectives, and build and share dashboards.

3.3 DATAIKU

Dataiku aka Dataiku Data Science Studio (DSS) provides a centralized data platform that moves businesses from scale and traditional analytics to Enterprise AI in their data journeys,



empowering self-service analytics while also enabling the operationalization of machine learning models in digital environments. Dataiku, the same as Informatica or Microsoft SSIS, is mainly a visual tool where you work with various components that connect to a complex data flow. Dataiku calls these transformations recipes. Some features of Dataiku and why we need to use it:

- 🚦 **End to end:** Dataiku DSS addresses the whole process of designing, deploying and operating a Data project from a single solution. It includes data preparation and cleansing, data exploration, data visualization, machine learning / predictive modeling, information delivery, automation, monitoring and API deployments.
- 🚦 **Collaboration:** Dataiku DSS is a web-based, multi-user environment where entire teams work together in a shared environment. It enables collaboration by providing features such as discussion/notifications, to-do lists, documentation, code and best practice sharing etc.
- 🚦 **On-Cloud:** Dataiku DSS can be implemented in public clouds and integrates with the various cloud providers.
- 🚦 **Multi-profile and inclusive:** Dataiku DSS is designed to be accessible (easy to adopt) and attractive to users with different profiles, from non-technical business analysts, all the way to top-flight data scientists and coders.
- 🚦 **Data agnostic:** Dataiku DSS natively connects to the full spectrum of technologies within a data ecosystem and leverages them to their highest potential.

- 🚦 **Open and compliant with IT:** Dataiku DSS is compliant with the IT standard in terms of security, and integration. The solution can be managed through its UI but also via Public API.

3.3.1 DATA, DATASETS & RECIPES

Dataiku DSS allows us to work with data that is structured or unstructured. Structured data is a series of records with the same schema. In Dataiku DSS, such data is referred to as a dataset. Unstructured data can have an internal structure, but the entries do not necessarily have the same schema. Examples of unstructured data are images, video, audio, and so forth.

The dataset is the core object we will be manipulating in DSS. It is analogous to a SQL table, as it consists of a series of records with the same schema. DSS supports various kinds of datasets. For example:

- 🚦 A SQL table or custom SQL query
- 🚦 A collection in MongoDB
- 🚦 A folder with data files on your server
- 🚦 A folder with data files on a Hadoop cluster.

Recipes are the building blocks of our data applications. Each time we make a transformation, an aggregation, a join, etc. with DSS, we will be creating a recipe. Recipes have input datasets and output datasets, and they indicate how to create the output datasets from the input datasets. Data Science Studio supports various kind of recipes:

- 🚦 Executing a data preparation script defined visually within the Studio
- 🚦 Executing a SQL query
- 🚦 Executing a Python script (with or without the use of the Pandas library)
- 🚦 Executing a Pig script
- 🚦 Executing a Hive query
- 🚦 Synchronizing the content of input to output datasets

Recipes and datasets together create the graph of the relationships between the datasets and how to build them. This graph is called the Flow. It is used by the dependencies management engine to automatically keep our output datasets up to date each time our input datasets or recipes are modified.

3.4 ROLES & RESPONSIBILITIES OF A DATA SCIENTIST

- ✚ **Management:** The Data Scientist plays an insignificant managerial role where he/she supports the construction of the base of futuristic and technical abilities within the Data and Analytics field in order to assist various planned and continuing data analytics projects.
- ✚ **Analytics:** The Data Scientist represents a scientific role where he plans, implements, and assesses high-level statistical models and strategies for application in the business's most complex issues. The Data Scientist develops econometric and statistical models for various problems including projections, classification, clustering, pattern analysis, sampling, simulations, and so forth.
- ✚ **Strategy/Design:** The Data Scientist performs a vital role in the advancement of innovative strategies to understand the business's consumer trends and management as well as ways to solve difficult business problems, for instance, the optimization of product fulfillment and entire profit.
- ✚ **Collaboration:** The role of the Data Scientist is not a solitary role and, in this position, he collaborates with superior data scientists to communicate obstacles and findings to relevant stakeholders in an effort to enhance drive business performance and decision-making.
- ✚ **Knowledge:** The Data Scientist also takes leadership to explore different technologies and tools with the vision of creating innovative data-driven insights for the business at the most agile pace feasible. In this situation, the Data Scientist also uses initiative in assessing and utilizing new and enhanced data science methods for the business, which he delivers to senior management of approval.
- ✚ **Other Duties:** A Data Scientist also performs related tasks and tasks as assigned by the Senior Data Scientist, Head of Data Science, Chief Data Officer, or the Employer.

3.4.1 DIFFERENCE BETWEEN DATA SCIENTIST, DATA ANALYST, AND DATA ENGINEER

Table 1: Difference between DS, DA and DE

Data Scientist	Data Analyst	Data Engineer
The focus will be on the futuristic display of data.	The main focus of a data analyst is on optimization of scenarios, for example how an employee can enhance the company's product growth.	Data Engineers focus on optimization techniques and the construction of data in a conventional manner. The purpose of a data engineer is continuously advancing data consumption.
Data scientists present both supervised and unsupervised learning of data, say regression and classification of data, Neural networks, etc.	Data formation and cleaning of raw data, interpreting and visualization of data to perform the analysis and to perform the technical summary of data.	Frequently data engineers operate at the back end. Optimized machine learning algorithms were used for keeping data and making data to be prepared most accurately.
Skills required for Data Scientist are Python, R, SQL, Pig, SAS, Apache Hadoop, Java, Perl, and Spark.	Skills required for Data Analyst are Python, R, SQL, and SAS.	Skills required for Data Engineer are MapReduce, Hive, Pig Hadoop, techniques.

Chapter 4

Conclusion

4.1 Conclusion

I am grateful for the internship at Deloitte – USI, which has provided me with invaluable corporate and technical expertise and experience. Technologies such as Data Science and Data Engineering are in-demand skillsets in today's world, and the need for this expertise is increasing. Data is the new oil. This statement demonstrates how data capture, storage, and analysis are at the heart of every contemporary IT system as well as major multi-national corporations. The thesis goes through Azure Data Factory, Azure Databricks, and Dataiku, which are utilized for Data Warehousing, Data Lake, and Data Lakehousing, respectively. The core concepts of Data Science and Data Engineering are used in DW, Data Lake, and DL.

4.2 Future Scope

As discussed in section 2.1, there are several applications of Data Science, and the scope and applications will only increase thereafter. Artificial Intelligence, the Internet of Things, and Deep Learning are just a few of the cutting-edge technologies covered by Data Science. Data science's influence has grown dramatically as a result of its success and technical advancements. The value of acquiring and collecting data is critical because it allows businesses to predict and affect our purchase behaviors. As a result, it wields significant power through its purchasing power.

Everyone generates data on a daily basis, both with and without our knowledge. As time goes on, the amount of data we engage with on a daily basis will only increase. Furthermore, the amount of data available on the planet will grow at breakneck pace. As data production grows, data scientists will be in high demand to assist businesses in effectively using and managing it.

References

- [1] Westin, Soffi & Stendal, Karen. (2018). “IMPLEMENTING CLOUD BASED BIG DATA PLATFORMS—A CASE USING MICROSOFT AZURE”, Big Data - Data Quality and Governance, NOKOBI, Svalbard
- [2] Milind Jadhav, Amol Goje, & Jitendra Chavan. (2021) “Data Migration From On Premise OracleDatabaseTo SQL Manage Instance On Azure Cloud Using Azure Data Factory -A Working Approach” – Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 3524-3528
- [3] Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, and Matei Zaharia. 2015. Spark SQL: Relational Data Processing in Spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15). Association for Computing Machinery, New York, NY, USA, 1383–1394. <https://doi.org/10.1145/2723372.2742797>
- [4] Shivaram Venkataraman, Zongheng Yang, Davies Liu, Eric Liang, Hossein Falaki, Xiangrui Meng, Reynold Xin, Ali Ghodsi, Michael Franklin, Ion Stoica, and Matei Zaharia. 2016. SparkR: Scaling R Programs with Spark. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16). Association for Computing Machinery, New York, NY, USA, 1099–1104. <https://doi.org/10.1145/2882903.2903740>
- [5] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. 2016. MLlib: machine learning in apache spark. J. Mach. Learn. Res. 17, 1 (January 2016), 1235–1241.
- [6] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: a unified engine for big data processing. Commun. ACM 59, 11 (November 2016), 56–65. <https://doi.org/10.1145/2934664>
- [7] Zaharia, M., Chowdhury, N., Franklin, M., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets [White paper]. EECS Department, University of California, Berkeley.
- [8] Michael Armbrust, Tathagata Das, Aaron Davidson, Ali Ghodsi, Andrew Or, Josh Rosen, Ion Stoica, Patrick Wendell, Reynold Xin, and Matei Zaharia. 2015. Scaling spark in the real world: performance and usability. Proc. VLDB Endow. 8, 12 (August 2015), 1840–1843. <https://doi.org/10.14778/2824032.2824080>
- [9] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica,

- and Matei Zaharia. 2010. A view of cloud computing. *Commun. ACM* 53, 4 (April 2010), 50–58. <https://doi.org/10.1145/1721654.1721672>
- [10] Herrera, C., Medina-Cetina, Z., Pompelli, G., Cochran, M., Olivares, M. & Perez-Patron, M (2021, January 1). CBTS-SGL Webinar - Breaking Silos. The power of collaboration and abstraction - Catalina Herrera (Dataiku) [Video] . <https://r13-cbts-sgl.engr.tamu.edu/cbts-sgl-webinar-breaking-silos-the-power-of-collaboration-and-abstraction-catalina-herrera-dataiku/>
- [11] Project Tungsten. <https://databricks.com/blog/2015/04/28/>
- [12] Kai Zeng, Sameer Agarwal, Ankur Dave, Michael Armbrust, and Ion Stoica. 2015. G-OLA: Generalized On-Line Aggregation for Interactive Analysis on Big Data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15)*. Association for Computing Machinery, New York, NY, USA, 913–918. <https://doi.org/10.1145/2723372.2735381>
- [13] Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, and Ion Stoica. 2014. GraphX: graph processing in a distributed dataflow framework. In *Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation (OSDI'14)*. USENIX Association, USA, 599–613.
- [14] Frank Austin Nothaft, Matt Massie, Timothy Danford, Zhao Zhang, Uri Laserson, Carl Yeksigian, Jey Kottalam, Arun Ahuja, Jeff Hammerbacher, Michael Linderman, Michael J. Franklin, Anthony D. Joseph, and David A. Patterson. 2015. Rethinking Data-Intensive Science Using Scalable Analytics Systems. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15)*. Association for Computing Machinery, New York, NY, USA, 631–646. <https://doi.org/10.1145/2723372.2742787>
- [15] Freeman, J., Vladimirov, N., Kawashima, T., Mu, Y., Sofroniew, N. J., Bennett, D. V., Rosen, J., Yang, C. T., Looger, L. L., & Ahrens, M. B. (2014). Mapping brain activity at scale with cluster computing. *Nature methods*, 11(9), 941–950. <https://doi.org/10.1038/nmeth.3041>
- [16] Azure Data Factory Documentation - Azure Data Factory. (n.d.). Microsoft Docs. Retrieved June 9, 2022, from <https://docs.microsoft.com/en-us/azure/data-factory/>
- [17] T. (n.d.-b). Python - Data Science Tutorial. Learn Python Data Science - Absolute Beginners. Retrieved June 9, 2022, from https://www.tutorialspoint.com/python_data_science/index.htm
- [18] D.D.S.S. (n.d.-a). Dataiku DSS — Dataiku DSS 10.0 documentation. Dataiku DSS. Retrieved June 9, 2022, from <https://doc.dataiku.com/dss/latest/>

- [19] D.D.S.S. (n.d.-b). DSS concepts — Dataiku DSS 10.0 documentation. DSS Concepts - Data, Datasets and Recipes. Retrieved June 9, 2022, from <https://doc.dataiku.com/dss/latest/concepts/index.html>
- [20] D. (2022b, June 9). Databricks documentation | Databricks on Azure. Databricks Documentation. Retrieved June 9, 2022, from <https://docs.databricks.com/index.html>
- [21] D. (2022, June 7). What is the Databricks Lakehouse? | Databricks on Azure. What Is the Databricks Lakehouse? Retrieved June 9, 2022, from <https://docs.databricks.com/lakehouse/index.html>
- [22] M.A. (n.d.-d). Azure Databricks documentation. Microsoft Docs. Retrieved June 9, 2022, from <https://docs.microsoft.com/en-us/azure/databricks/>
- [23] KOMTAŞ Bilgi Yönetimi: Veri Yönetimi ve Analitik Uygulamalardaki Çözüm Ortağınız. (n.d.). KOMTAS Dataiku © 2021. Retrieved June 9, 2022, from <https://komtas.com/technology-detail/dataiku/why-use-dataiku>
- [24] Wikipedia contributors. (2022, April 12). Data warehouse. Wikipedia. Retrieved June 9, 2022, from https://en.wikipedia.org/wiki/Data_warehouse
- [25] Wikipedia contributors. (2022b, April 21). Data lake. Wikipedia. Retrieved June 9, 2022, from https://en.wikipedia.org/wiki/Data_lake
- [26] Kutay, J. (2022, April 20). Data Warehouse vs. Data Lake vs. Data Lakehouse: An Overview of Three Cloud Data Storage Patterns. Striim. Retrieved June 9, 2022, from <https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview/>
- [27] Arora, S. K. (n.d.). 10 Top Data Science Applications in 2022 [Updated]. Hackr.io. Retrieved June 9, 2022, from <https://hackr.io/blog/top-data-science-applications>
- [28] GeeksforGeeks. (2021, March 12). Major Applications of Data Science. Retrieved June 9, 2022, from <https://www.geeksforgeeks.org/major-applications-of-data-science/>
- [29] GeeksforGeeks. (2020, December 8). What Are the Roles and Responsibilities of a Data Scientist? Retrieved June 9, 2022, from <https://www.geeksforgeeks.org/what-are-the-roles-and-responsibilities-of-a-data-scientist/>
- [30] E. (2022c, April 5). What is the Future Scope of Data Science? Edureka. Retrieved June 9, 2022, from <https://www.edureka.co/blog/future-scope-of-data-science/>
- [31] J. (2022d, January 15). Create an Azure data factory using the Azure Data Factory UI - Azure Data Factory. Microsoft Docs. Retrieved June 9, 2022, from

<https://docs.microsoft.com/en-us/azure/data-factory/quickstart-create-data-factory-portal>

- [32] M.D. (n.d.-e). Azure Databricks. Azure Databricks Design AI with Apache Spark™-Based Analytics. Retrieved June 9, 2022, from <https://azure.microsoft.com/en-us/services/databricks/#overview>
- [33] S. (2022e, April 18). Data Scientist Job Description: Role, Responsibilities and Skills Required. Simplilearn.Com. Retrieved June 9, 2022, from <https://www.simplilearn.com/data-scientist-job-description-article>
- [34] Bansal, S. (2021, December 2). Future Scope of Data Science - Career in Data Science. Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. Retrieved June 9, 2022, from https://www.analytixlabs.co.in/blog/scope-of-data-science/#Future_Scopes_of_Data_Science
- [35] Wikipedia contributors. (2022c, June 6). Deloitte. Wikipedia. Retrieved June 9, 2022, from <https://en.wikipedia.org/wiki/Deloitte>
- [36] Women in Data Science and Analytics. (n.d.). Deloitte United States. Retrieved June 9, 2022, from <https://www2.deloitte.com/us/en/pages/about-deloitte/solutions/women-in-data-science.html>