Customized Data Engineering Roadmap

Project Assignment - 01

Aurabrato Ghosh

Azure Data Engineering

Table of Contents

Review of Week 1

Cloud Computing and Azure Architecture

Cloud computing provides virtualized computing resources over the internet, eliminating physical hardware management. Key benefits include scalability, pay-as-you-go pricing, and global accessibility. The major cloud service providers are Microsoft Azure, AWS, and Google Cloud Platform.

The logical architecture is organized into:

- **Tenant:** Represents the organization within Azure.
- **Subscription:** Billing and resource management unit.
- **Resource Groups:** Logical containers for managing related Azure resources.
- **Resources:** The actual cloud services like databases and storage accounts.

Azure Subscription Types

- **Free Trial:** Includes initial credits for 30 days.
- **Student Access:** Provides credits for a full year.
- **Pay-as-you-go:** Charges based on actual usage.

Tenant IDs uniquely identify an organization's Azure environment, and in corporate settings,

Region Selection and Data Management

- Selecting the appropriate region affects latency, data compliance, and costs.
- This is important for real-time applications to optimize performance.

Types of Data

- **Structured Data:** Highly organized, stored in rows and columns (e.g., SQL databases).
- **Semi-structured Data:** Some organizational elements like JSON, XML.
- **Unstructured Data:** No predefined format, e.g., images, videos.

Parquet File Format

The Parquet File Format is a columnar storage format optimized for big data tools. It provides,

- Efficient storage and retrieval.
- High compression and reduced costs.
- Optimized queries by scanning only relevant columns.

Azure Storage Accounts and Types

Azure Storage is the main container for storing various data types, with two primary storage models:

- **General Purpose (Blob Storage):** Used for unstructured data like images and backups.
- **Azure Data Lake Storage Gen2 (ADLS Gen2):** Optimized for big data analytics, with hierarchical namespaces and faster processing.

Advantages of ADLS gen2 Storage account

- **Hierarchical Namespace:** Allows folder and directory structuring.
- **Access Control Lists (ACLs):** Granular permission settings.
- **Serverless Pool Capabilities:** Enables SQL-based queries directly on stored data.

Storage Account Access Tiers

- **Hot:** Frequently accessed data.
- **Cool:** Infrequently accessed data.
- **Cold:** Low-cost storage for archived data.
- **Archive:** Cheapest.

Role-Based Access Control (RBAC)

- **Storage Blob Data Contributor:** For uploading, downloading, and modifying data.
- **Storage Blob Data Reader:** For read-only access.
- **Data Owner:** For full control over the data environment.

Review of Week 2

Azure Storage Management & Cost Optimization

We began by comparing GRS (Geo-Redundant Storage) and LRS (Locally Redundant Storage). GRS offers high availability across regions but comes at nearly double the cost of LRS, which is sufficient for non-critical workloads. The Soft Delete feature in Blob storage was introduced for recovering deleted files with configurable retention up to 365 days.

ETL vs. ELT in Cloud Workflows

Learnt the relevance of ETL in legacy tools like SSIS, versus ELT in modern cloud tools like ADF and Databricks. ELT is preferred in cloud due to scalability and processing efficiency. ADLS Gen2 is best for hierarchical, structured data, while Blob Storage is ideal for flat, archival storage.

ADF Interface and Components

We worked hands-on in Azure Data Factory, exploring the Author, Monitor, and Manage tabs.

- Pipelines for orchestration.
- Activities like Copy Data.
- Linked Services for connection credentials.

Pipeline Execution, Debugging & Validation

We used the Debug function to test pipelines and publish to deploy. It was important to validate that the source files remained unchanged post-transformation and that the expected output appeared in the destination container.

Git Integration & Version Control

ADF supports Git integration, allowing version control. ADF objects like pipelines and datasets are stored as JSON files, enabling code-based tracking and rollback.

CSV File Merging

We merged multiple CSVs using wildcard folder paths, storing results in a new folder with "none" schema selected. Merging did not preserve order, and deduplication had to be handled separately.

LinkedIn Profile Analysis

Common Technical Skills

Most professionals had hands-on experience with tools like Azure Data Factory, Azure Data Lake, and Azure Databricks. Python and SQL were almost always listed as core programming languages, often paired with PySpark for large-scale data processing. Building and managing ETL pipelines was a common theme across almost all profiles.

Certifications

Many have earned Microsoft Certified: Azure Fundamentals and Azure Data Engineer Associate. Others also had Databricks and Cloud Practitioner certifications.

Educational Background

Most individuals came from engineering or computer science backgrounds, holding either a Bachelor's or Master's degree. A few had transitioned from non-traditional academic paths.

Career Progression

It was interesting to see that many data engineers began in roles like ETL Developer, BI Analyst, or Programmer Analyst. Over time, they transitioned into more advanced roles such as Senior Data Engineer, Azure Architect, or even Cloud Solution Architect.

Job Posting Analysis

Role Expectations

The job titles I came across ranged from Junior Data Engineer and Azure System Analyst to Cloud Architect and Senior Data Engineer. Some roles were very cloud-specific (e.g., Azure Cloud Engineer), while others had a broader focus, incorporating elements of DevOps and platform optimization.

Organizations Hiring Data Engineers

These roles were offered by a mix of companies—from large consulting firms like EY to utilities companies like Canadian Utilities Ltd., and tech-driven startups such as Bree and Lumenalta. This mix showed me that the demand for data engineers is not limited to the tech industry.

Required Experience Levels

Job postings were generally grouped into three categories: entry-level (0–2 years), mid-level (3–5 years), and senior (5+ years).

- Entry-level roles expected basic cloud exposure and strong coding fundamentals.
- Mid-level positions asked for proven experience in building and managing cloud-based data pipelines.
- Senior positions demanded end-to-end system design, architecture planning, and leadership in cloud migration strategies.

Skills and Tools Required

The most frequently listed tools and technologies were Azure Data Factory, Databricks, and scripting languages such as Python. Knowledge of CI/CD pipelines, DevOps practices, and data pipeline orchestration tools was often listed as mandatory for mid and senior roles. Database technologies like PostgreSQL, SQL Server, and NoSQL systems were commonly expected.

Salary Compensation

- Entry-level positions offered between CAD 60,000 to 90,000.
- Mid-level roles offered CAD 80,000 to 120,000.
- Senior roles ranged up to CAD 190,000 annually.

Self-Assessment

Academic & Technical Background

My coursework has introduced me to core concepts in systems analysis, data processing, and machine learning. I also have professional experience working as a Systems Engineer at Infosys, where I was involved in mainframe-based enterprise systems, which helped me develop logical thinking and an appreciation for large-scale data infrastructure.

Strengths

- Familiar with programming languages such as Python and SQL.
- Completed the Microsoft Azure Data Fundamentals learning path and currently exploring Azure services in-depth for the certification.
- I earned an Agile certification and received formal training in Jira and Confluence, which has enhanced my understanding of agile team collaboration, task tracking, and sprint management.

Areas for Improvement

- Limited practical exposure to Azure services like Data Factory, Databricks, and Synapse Analytics - which I am actively working on.
- I need to deepen my knowledge of big data frameworks, especially Apache Spark and PySpark.
- Hands-on experience in building automated ETL pipelines and deploying them.

Current Focus

I'm working on strengthening my command of core tools such as SQL and Python. My next focus is building confidence in Azure services and gaining certification (DP-900, DP-203).

Roadmap Development

Skill Development Plan

Skill / Tool	Current Level	Desired Proficiency
SQL	7	10
Python	6	9
Azure Data Factory	1	10
Azure Databricks	0	10
Azure Synapse Analytics	0	10
PySpark	1	9
Power BI / Visualization	2	7
Agile Methodology	6 (Certified)	8
Jira & Confluence	6 (Certified)	8

Certifications Roadmap

- Short-Term Goals:
  - DP-900: Microsoft Azure Data Fundamentals.
- Long-Term Goals:
  - DP-203: Azure Data Engineer Associate.
  - Databricks Lakehouse Fundamentals.
  - Fabric Data Engineer.

Learning and Execution Phases

- **Refine Core Skills:** Finalize SQL and Python practice with projects and case studies.
- **Master Azure Basics:** Deepen understanding of Azure portal, subscriptions, and storage types.
- **Practice Tools:** Learn Azure Data Factory, Databricks, and Synapse.
- **Explore Big Data & Spark:** Study distributed computing, Spark APIs, and PySpark workflows.
- **Implement Projects:** Build end-to-end pipelines (batch and streaming) using Azure services.
- **Work in Agile Setup:** Use Jira and Confluence to simulate agile data projects and sprint tasks.
- **Build Portfolio:** Document learning and projects on GitHub with clean documentation.

My goal is to become an Azure-focused data engineer capable of designing scalable, secure, and automated data workflows in the cloud.