Student's Full Name: Surajit Pal

Course Title: Data Warehousing and Analytics in the Cloud

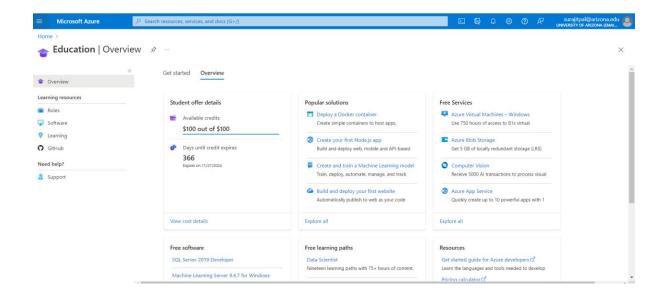
Term name and year: Spring 2023

**Submission Week: Week 6 – Assignment 4** 

Instructor's Name: Dr.Nayem Rahman

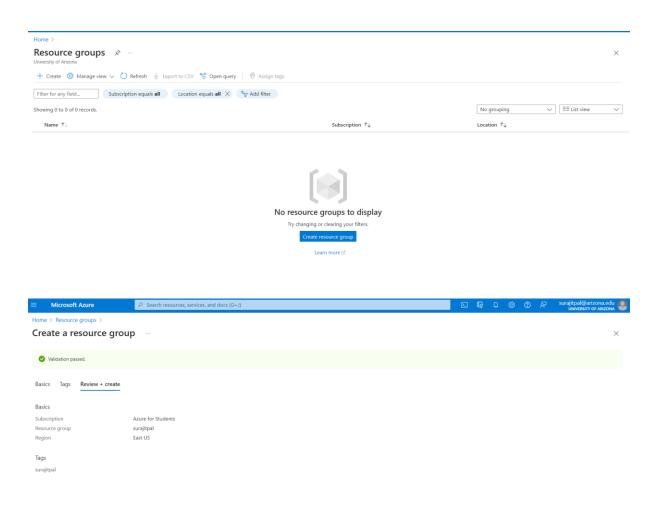
**Date of Submission: 28-Nov-2023** 

## Q1. Create your Azure free account

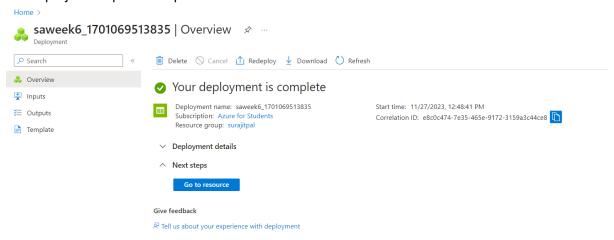


An Azure account has been set up with Arizona credentials, allowing the user to possess and utilize resources on the Azure Cloud.

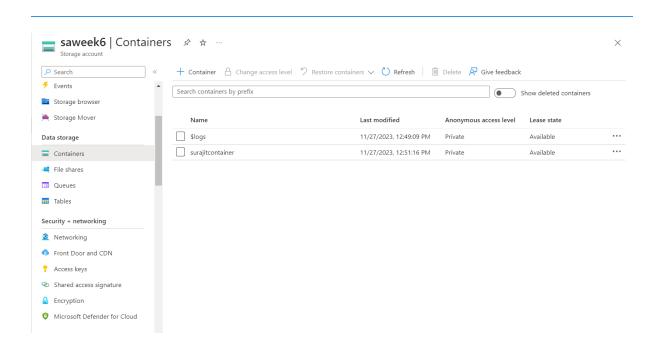
## Q2. Create a container in Azure Storage, and to upload and download block blobs in that container.

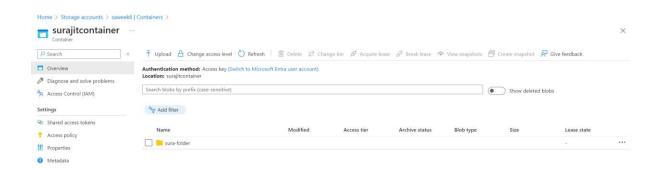


The first step involves creating a Resource Group to organize associated resources together according to the project or specific scope of work.

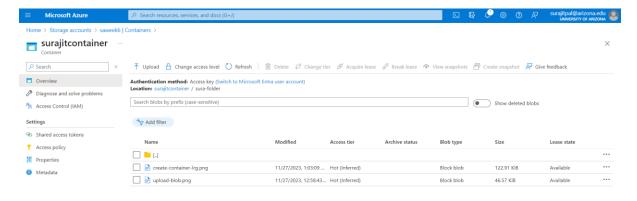


The second step is to establish a Storage Account, which is necessary for accessing and managing data storage, accommodating various data formats like RDBMS, documents, CSV files, and others.

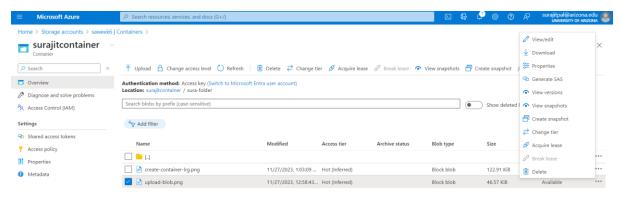




Under the storage account, a container is created to organize data into separate collections, and within it, a folder is set up for uploading Blobs. Blob storage is specifically optimized for holding vast amounts of unstructured data.

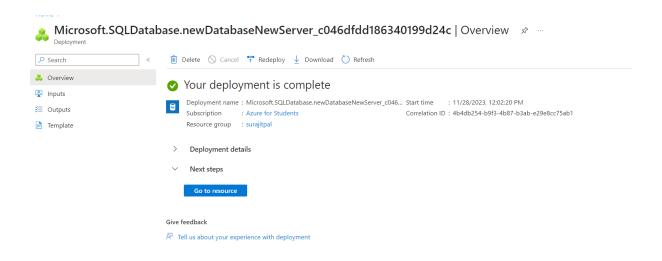


The 'Upload block blobs' feature is utilized for uploading files in bulk, which can include various file types such as plain text, documents, graphics, and other formats.

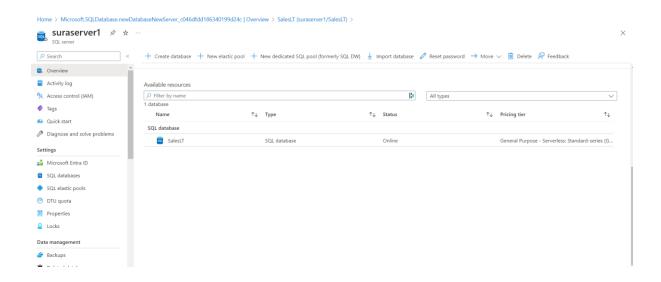


Files can be downloaded by clicking on the three-dot button next to the selected files.

## Q3. Create a SQL database on Azure



To create a SalesLT database with sample data, initiate the database creation workflow which will prompt you for configuration inputs such as Resource Group, Server name, database name, Network settings, Authentication, and other relevant details



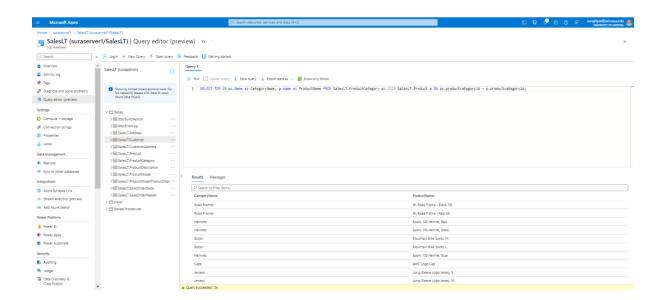
### Q4. Summarize SQL database creation in Azure

Creating a SQL database in Azure requires a sequence of steps:

- 1. Start by establishing a resource group.
- 2. Then, set up a blob storage account.
- 3. Use the Azure graphical user interface (GUI) to create the SQL database.
- 4. During this process, you'll need to create a hosting server and configure settings such as performance tiers and security features.
- 5. With these steps complete, your SQL database will be operational.

Through this process, we will discover Azure SQL Database's scalability options, like choosing the right performance tier to match your application's needs, and you'll learn about the security measures available, including firewall settings and authentication protocols, to protect against unauthorized access.

### Q5. Query the SQL database in Azure



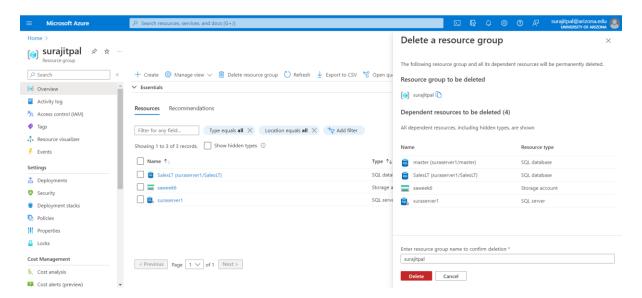
To retrieve the top 10 rows of product and category data from a database, execute an SQL query that selects the first 10 rows from the relevant tables.

### Q6. Describe the SQL database "SalesLT" on Azure.

The "SalesLT" database in Azure comprises 10 tables, each with the columns "rowguid" and "ModifiedDate" to maintain an audit trail of record modifications. Below is a summary of each table and its primary key (PK):

- 1. Address (PK: AddressID): Holds address information.
- 2. Customer (PK: CustomerID): Contains customer information.
- 3. CustomerAddress (PK: CustomerID, AddressID): A junction table linking Customers to Addresses, clarifying many-to-many relationships.
- 4. Product (PK: ProductID): Stores product information.
- 5. ProductCategory (PK: ProductCategoryID): Details categories of products.
- 6. ProductDescription (PK: ProductDescriptionID): Describes products.
- 7. ProductModel (PK: ProductModelID): Provides information on product models.
- 8. ProductModelProductDescription (PK: ProductModelID, ProductDescriptionID, Culture): Another junction table associating Product Models with their Descriptions, featuring a composite primary key that includes a non-numeric "Culture" column.
- 9. SalesOrderDetail (PK: SalesOrderID, SalesOrderDetailID): Records details for each product in a sales order.
- 10. SalesOrderHeader (PK: SalesOrderID): Captures high-level sales order information, such as customer and total order value.

## Q7. Clean up resources in Azure



In Azure, to remove all resources and objects within a certain scope, use the "Delete a resource group" feature. This action will delete everything contained in the resource group, including containers, servers, databases, and any other associated resources, effectively cleaning up your Azure environment.

Q8. Summarize the article on "Understanding Data Store Models" . Link --> <a href="https://learn.microsoft.com/en-us/azure/architecture/guide/technology-choices/data-store-overview">https://learn.microsoft.com/en-us/azure/architecture/guide/technology-choices/data-store-overview</a>

The article "Understanding Data Store Models" from Microsoft's Azure Architecture Guide provides a comprehensive overview of various data store technologies and their applications. Here are the key takeaways:

- Polyglot Persistence and Storage Models: The concept of polyglot persistence is introduced, emphasizing the use of multiple data store technologies to handle heterogeneous data effectively. The article explains that different types of data are best stored in different data stores, each optimized for specific workloads or usage patterns.
- Relational Database Management Systems (RDBMS): RDBMS are described as systems
  organizing data in two-dimensional tables with rows and columns, typically using SQL. They
  are suited for scenarios requiring strong consistency guarantees and atomic transactions but
  face limitations in horizontal scalability.
- Key/Value Stores: These stores associate each data value with a unique key, optimized for simple lookups. They are highly scalable and suitable for applications like data caching and session management but less effective for complex queries across different key/value pairs.
- Document Databases: Document databases store data in documents containing named fields and data, which can be simple values or complex elements. They are flexible in terms of schema and are well-suited for applications where data structures align with object structures in application code.
- Graph Databases: Graph databases store data as nodes and edges, efficiently handling complex queries and analyses of relationships between entities. They are ideal for use cases like social graphs, organization charts, and recommendation engines.
- Other Data Store Models: The article also touches on other models like column-family databases, search engine databases, time series databases, and object storage, each with unique characteristics and suited for specific types of workloads and data.
- Choosing the Right Data Store: The article advises considering the storage model best suited
  for the application's requirements first, then selecting a specific data store within that
  category based on factors like feature set, cost, and ease of management.

This overview provides a foundational understanding of different data storage models and their appropriate use cases, aiding in the selection of the right data store technology for specific needs.

# Q9. Summarize the article on "Understanding the differences between NoSQL and relational databases". Link → <a href="https://learn.microsoft.com/en-us/azure/cosmos-db/relational-nosql">https://learn.microsoft.com/en-us/azure/cosmos-db/relational-nosql</a>

The article "Understanding Distributed NoSQL Databases" from Microsoft Azure focuses on the differences between NoSQL and relational databases, particularly in the context of Azure Cosmos DB. Here are the key takeaways:

- Challenges with Traditional Database Systems: The article begins by highlighting the
  challenges faced by traditional database systems, particularly those that enforce strict ACID
  semantics. While these systems ensure high consistency, they often struggle with
  concurrency, latency, and availability due to architectural restrictions. This leads to complex
  workarounds like manual data distribution or sharding.
- Introduction to NoSQL Databases: NoSQL databases are presented as a solution to the limitations of traditional databases. They are designed to simplify horizontal scaling and offer configurable levels of consistency. This flexibility allows NoSQL databases to scale across many nodes, offering speed and availability that can be better aligned with application needs.
- Distributed Databases: The article distinguishes between NoSQL databases and distributed databases. While many NoSQL databases are built for scalability, not all are inherently distributed. Setting up a NoSQL database for local or global distribution can be complex, involving significant planning and networking.
- Azure Cosmos DB: Azure Cosmos DB is highlighted as a platform that combines the benefits
  of distributed databases with the flexibility of NoSQL. It offers distributed data APIs for both
  NoSQL and relational models. Azure Cosmos DB allows for fine-tuning of consistency and
  availability to meet specific application requirements. It can automatically distribute data
  locally or globally and provides ACID guarantees while scaling throughput according to
  application needs.
- Next Steps and Additional Resources: The article concludes by suggesting further exploration into distributed relational databases and provides resources for getting started with various APIs in Azure Cosmos DB, including NoSQL, MongoDB, Cassandra, Gremlin, and Table.

In summary, the article provides an overview of the challenges with traditional database systems, introduces NoSQL and distributed databases, and explains how Azure Cosmos DB addresses these challenges by offering a flexible, scalable, and globally distributed database solution.

## Q10. Summarize the article on "Understanding Azure Cosmos DB". Link --> https://learn.microsoft.com/en-us/azure/cosmos-db/introduction

The article "Understanding Azure Cosmos DB" provides a detailed overview of Azure Cosmos DB, a fully managed NoSQL and relational database service designed for modern application development. Here are the key takeaways:

- High Responsiveness and Availability: Azure Cosmos DB is tailored for applications that need
  to be highly responsive and always online. It achieves low latency and high availability by
  deploying in datacenters close to users, essential for applications that must respond in real
  time to large changes in usage.
- Unified AI Database: Azure Cosmos DB serves as a single database solution for various operational data models, including relational, document, vector, key-value, graph, and table.
   This makes it particularly suitable for AI-powered applications that integrate multiple data stores.

### **Key Benefits:**

- Guaranteed Speed at Any Scale: Azure Cosmos DB offers single-digit millisecond response times and automatic scalability, with speed and throughput backed by SLAs.
- b. Simplified Application Development: It supports open-source APIs and multiple SDKs, allowing for schemaless data and no-ETL analytics over operational data. It's deeply integrated with key Azure services and offers various database APIs.
- c. Mission-Critical Readiness: The service guarantees business continuity with 99.999% availability and provides enterprise-level security.
- d. Fully Managed and Cost-Effective: Azure Cosmos DB is a fully managed service, offering automatic maintenance, patching, and updates. It provides cost-effective options for various workloads, including a serverless model for managing traffic bursts.
- Azure Synapse Link for Azure Cosmos DB: This feature enables near real-time analytics over operational data, integrating tightly with Azure Synapse Analytics for reduced complexity and optimized large-scale analytics workloads.
- Versatility Beyond AI Database: Azure Cosmos DB is also ideal for web, mobile, gaming, and IoT applications, handling massive amounts of data with high throughput, low latency, and tunable consistency.
- Getting Started and Resources: The article provides resources for getting started with Azure Cosmos DB for various APIs (NoSQL, MongoDB, Cassandra, Gremlin, Table, PostgreSQL) and guides on choosing the right API based on workload requirements.

In summary, Azure Cosmos DB is a versatile, high-performance database service that caters to a wide range of modern application needs, from AI to IoT, offering scalability, low latency, and seamless integration with other Azure services.

# Q11. Summarize the article on "Azure Cosmos DB API for MongoDB". Link --> https://learn.microsoft.com/en-us/azure/cosmos-db/mongodb/mongodb-introduction

The article "Azure Cosmos DB API for MongoDB" provides an overview of Azure Cosmos DB's compatibility and benefits when used as a MongoDB database. Here are the key points:

- Compatibility with MongoDB: Azure Cosmos DB for MongoDB allows users to use Azure
  Cosmos DB as if it were a MongoDB database. It supports existing MongoDB skills, drivers,
  SDKs, and tools, enabling easy integration with existing MongoDB applications.
- Benefits of Cosmos DB for MongoDB:
  - a) Request Unit (RU) Architecture: Offers flexible scaling with Request Units, instantaneous scalability, automatic and transparent sharding, and 99.999% availability. It supports active-active database configurations across multiple regions, unlike MongoDB Atlas which only supports active-passive deployments.
  - Real-Time Analytics (HTAP) at Any Scale: Enables analytics on transactional MongoDB data in real time without affecting the database, using a cloud-native analytical columnar store.
  - c) Serverless Deployments: Offers a serverless capacity mode, charging per operation and eliminating costs when the database is not in use.
- vCore Architecture: This architecture provides dedicated instances for MongoDB apps, native vector search integration, flat pricing, text indexing for efficient querying, and highcapacity vertical scaling without the need for a shard key. It also includes free 35-day backups with point-in-time restore.
- How It Works: Azure Cosmos DB for MongoDB implements the MongoDB wire protocol, allowing compatibility with MongoDB client SDKs, drivers, and tools without hosting the MongoDB database engine itself.
- Important Note: The article clarifies that Azure Cosmos DB provides wire protocol compatibility with MongoDB databases but does not run MongoDB databases to provide this service. It is not affiliated with MongoDB, Inc.

In summary, Azure Cosmos DB for MongoDB offers a highly compatible, scalable, and feature-rich environment for MongoDB applications, with benefits like real-time analytics, serverless deployments, and a choice between RU and vCore architectures.

# Q12. Summarize the article on "Nodes and tables in Azure Database for PostgreSQL – Hyperscale". Link --> <a href="https://learn.microsoft.com/en-us/azure/postgresql/hyperscale/concepts-nodes">https://learn.microsoft.com/en-us/azure/postgresql/hyperscale/concepts-nodes</a>

The article "Nodes and Tables in Azure Database for PostgreSQL – Hyperscale" provides insights into the architecture and functioning of Azure Cosmos DB for PostgreSQL. Here are the key takeaways:

- Nodes in Azure Cosmos DB for PostgreSQL:
  - a) The system uses a "shared nothing" architecture with multiple PostgreSQL servers (nodes) coordinating in a cluster.
  - b) Each cluster has a coordinator node and multiple worker nodes. The coordinator node relays queries to the relevant workers and accumulates results.
  - c) The architecture allows the database to scale by adding more nodes, holding more data, and using more CPU cores than a single server.

### Table Types:

- a) Distributed Tables: These are horizontally partitioned across worker nodes. Rows of the table are stored on different nodes in shard tables.
- b) Reference Tables: Entire contents are concentrated into a single shard replicated on every worker. Useful for small data relevant to queries running on any worker node.
- c) Local Tables: Regular tables on the coordinator node that are not sharded. Suitable for small administrative tables.
- d) Local Managed Tables: Automatically added to metadata if a foreign key reference exists between a local table and a reference table. They can be queried from any node.
- e) Schema Tables: Introduced in Citus 12.0, these are associated with distributed schemas and automatically converted to colocated distributed tables without a shard key.

#### • Shards and Shard Placements:

- a) Distributed tables are stored as shards on worker nodes.
- b) The `pg\_dist\_shard` metadata table on the coordinator contains information about each shard.
- c) Shard placements determine which worker node holds a particular shard, and the coordinator node rewrites queries to target these specific shards.

### • Coordinator and Worker Roles:

- a) The coordinator node decides how to route queries (either to a single worker node or parallelized across several) based on the distribution of data.
- b) Schema-based sharding allows the coordinator to route queries directly to the node hosting the schema.

### • Distribution Column:

a) Azure Cosmos DB for PostgreSQL uses algorithmic sharding based on the value of a designated distribution column in a table.

### Next Steps:

a) The article suggests determining the application's type for data modeling and inspecting shards and placements with diagnostic queries.

In summary, Azure Cosmos DB for PostgreSQL offers a scalable, distributed database solution with various types of tables and an efficient system for managing data distribution and query routing across multiple nodes.

Q13. Summarize the article on "Overview - Azure Database for PostgreSQL - Flexible Server". Link --> <a href="https://learn.microsoft.com/en-us/azure/postgresql/flexible-server/overview">https://learn.microsoft.com/en-us/azure/postgresql/flexible-server/overview</a>

The article "Overview - Azure Database for PostgreSQL - Flexible Server" provides an in-depth look at the Flexible Server deployment model for Azure Database for PostgreSQL. Here are the key takeaways:

- Flexible Control and Customization: Azure Database for PostgreSQL Flexible Server is
  designed for granular control over database management and configurations, ideal for
  applications requiring specific customizations and high availability.
- High Availability and Automated Maintenance: The architecture supports high availability across multiple availability zones, with automated patching for hardware, OS, and PostgreSQL engine, and the ability to set custom maintenance windows.
- Performance and Scalability: Offers three compute tiers (Burstable, General Purpose, Memory Optimized) to suit various workloads, from development to high-performance production environments.
- Security and Backup Features: Includes enterprise-grade security with AES 256-bit encryption for data at rest and TLS 1.2 for data in motion. Automatic backups are provided with customizable retention periods.
- Cost Optimization and Global Availability: Enables cost-effective management with ondemand server start/stop capabilities and is available in multiple Azure regions globally.

In summary, Azure Database for PostgreSQL - Flexible Server provides a highly flexible, secure, and scalable database solution with advanced features for high availability, performance tuning, and cost management.

# Q14. Summarize the article on "Azure Database for PostgreSQL - Single Server". Link--> https://learn.microsoft.com/en-us/azure/postgresql/single-server/concepts-servers

The article "Azure Database for PostgreSQL - Single Server" provides an overview of the Single Server deployment option in Azure Database for PostgreSQL. Here are the key takeaways:

- Server Structure and Management: An Azure Database for PostgreSQL server in the Single Server model acts as a central administrative point for multiple databases. It is similar to the traditional PostgreSQL server, offering managed services with performance guarantees and server-level access and features. The server is created within an Azure subscription and is the parent resource for databases, providing a namespace and a connection endpoint for server and database access.
- Database Creation and Pricing: Within a single server, one can create multiple databases, either utilizing all resources for a single database or sharing resources among multiple databases. The pricing is structured per server, based on the configuration of the pricing tier, vCores, and storage.
- Authentication and Security: The server setup includes credentials for an admin user with
  the highest privilege, belonging to the role azure\_pg\_admin. However, this role does not
  have full superuser permissions, which are reserved for the azure\_superuser managed by
  the service. The server also includes default databases for maintenance and system
  processes.
- Server Parameters and Configuration: Azure Database for PostgreSQL allows viewing and editing a subset of PostgreSQL server parameters through the Azure portal or Azure CLI.
   These parameters are pre-configured with default values upon server creation, and some parameters requiring server restart or superuser access are not configurable by the user.
- Retirement Path and Migration: The article notes that Azure Database for PostgreSQL -Single Server is on the retirement path, and users are recommended to upgrade to Azure Database for PostgreSQL - Flexible Server. Resources are provided for migration and further understanding of the service.

In summary, Azure Database for PostgreSQL - Single Server offers a managed, scalable PostgreSQL server environment with customizable databases, pricing options, and security features, but it is moving towards retirement in favor of the Flexible Server model.

# Q15. Summarize the article on "What is Azure Database for PostgreSQL?". Link --> https://learn.microsoft.com/en-us/azure/postgresql/single-server/overview

The article "What is Azure Database for PostgreSQL?" provides an overview of Azure Database for PostgreSQL, focusing on its deployment models and key features. Here are the main points:

- Azure Database for PostgreSQL: This is a relational database service based on the PostgreSQL open-source relational database, offering built-in high availability, data protection with automatic backups, automated maintenance, predictable performance, elastic scaling, enterprise-grade security, and simplified management.
- Deployment Models: Azure Database for PostgreSQL is available in two deployment modes:
- Single Server: A fully managed database service optimized for built-in high availability, supporting PostgreSQL versions 9.5, 9.6, 10, and 11. It offers three pricing tiers (Basic, General Purpose, Memory Optimized) and is suited for cloud-native applications that can handle automated patching.
- Flexible Server: Designed for more granular control and flexibility over database management and configurations. It supports high availability across multiple zones, cost optimization controls, and is ideal for applications requiring more control and customization.
- Retirement Path: The article notes that Azure Database for PostgreSQL Single Server is on the retirement path, and users are encouraged to upgrade to the Flexible Server model.
- Global Availability: Both Single Server and Flexible Server models are available in a wide variety of Azure regions.
- Next Steps: The article suggests learning more about the deployment models to choose the right option based on specific needs.

In summary, Azure Database for PostgreSQL offers two main deployment models, Single Server and Flexible Server, each catering to different needs and offering a range of features from high availability and automated maintenance to flexible management and scalability. The Single Server model is being phased out in favor of the more flexible and customizable Flexible Server model.

# Q16. Summarize the article on "Azure Database for MySQL - Flexible Server". Link --> https://learn.microsoft.com/en-us/azure/mysql/flexible-server/overview

The article "Azure Database for MySQL - Flexible Server" provides a comprehensive overview of this fully managed database service. Here are the key takeaways:

- Flexible Server Deployment Model: Azure Database for MySQL Flexible Server is designed
  for granular control and flexibility in database management and configuration. It supports
  high availability within a single availability zone and across multiple zones, and offers cost
  optimization controls like the ability to stop/start servers and a burstable compute tier.
- Compute Tiers and Scalability: The service offers three compute tiers Burstable, General Purpose, and Business Critical - each catering to different workload requirements. Dynamic scalability allows databases to respond to changing resource needs, with users only paying for the resources they use.
- High Availability and Automated Maintenance: The service includes options for zone
  redundant and same-zone high availability, ensuring data protection and uptime. Automated
  patching for hardware, OS, and the MySQL engine is provided, with options for custom
  patching schedules.
- Security and Backup Features: Azure Database for MySQL Flexible Server employs AES 256bit encryption for data at rest and enforces TLS 1.2 for data in motion. It also offers automatic backups with customizable retention periods and encrypted storage.
- Additional Features: The service supports up to 10 read replicas for scaling read workloads, data-in replication for hybrid or multi-cloud synchronization, and enterprise-grade security and compliance features.

In summary, Azure Database for MySQL - Flexible Server is a versatile and scalable database service, offering a range of features for high availability, performance, security, and cost optimization, suitable for a variety of application development and deployment scenarios.

# Q17. Summarize the article on "Azure Database for MySQL Single Server". Link --> https://learn.microsoft.com/en-us/azure/mysql/single-server/single-server-overview

The article "Azure Database for MySQL Single Server" provides an overview of this deployment model in Azure Database for MySQL. Here are the key points:

- Overview and High Availability: Azure Database for MySQL Single Server is a fully managed database service optimized for minimal customization. It offers built-in high availability with 99.99% availability in a single zone, separating compute and storage for resilience. The service supports MySQL versions 5.6 (retired), 5.7, and 8.0 and is available in a wide variety of Azure regions.
- Automated Patching and Backups: The service includes automated patching of hardware,
  OS, and the MySQL engine, with no user action required. It also automatically creates server
  backups, stored in locally redundant or geo-redundant storage, with a default retention
  period of seven days, extendable to 35 days.
- Performance and Scalability: Single Server is available in three SKU tiers: Basic, General Purpose, and Memory Optimized, catering to different workload needs from low-cost development to high concurrency production workloads. It supports online storage scaling and storage autogrowth.
- Security and Compliance: The service uses FIPS 140-2 validated cryptographic module for storage encryption and enforces TLS for data in motion. It supports private access, threat protection, and Microsoft Entra ID authentication, and is compliant with industry certifications like FedRAMP, HIPAA, and PCI DSS.
- Migration Support and Monitoring: Azure Database for MySQL Single Server supports
  various migration methods, including Dump and Restore, Azure Database Migration Service,
  and data-in replication. It also features built-in performance monitoring and alerting, with
  tools like Query Store for workload optimization.

The article emphasizes that Azure Database for MySQL - Single Server is on the retirement path, recommending users to upgrade to Azure Database for MySQL - Flexible Server.