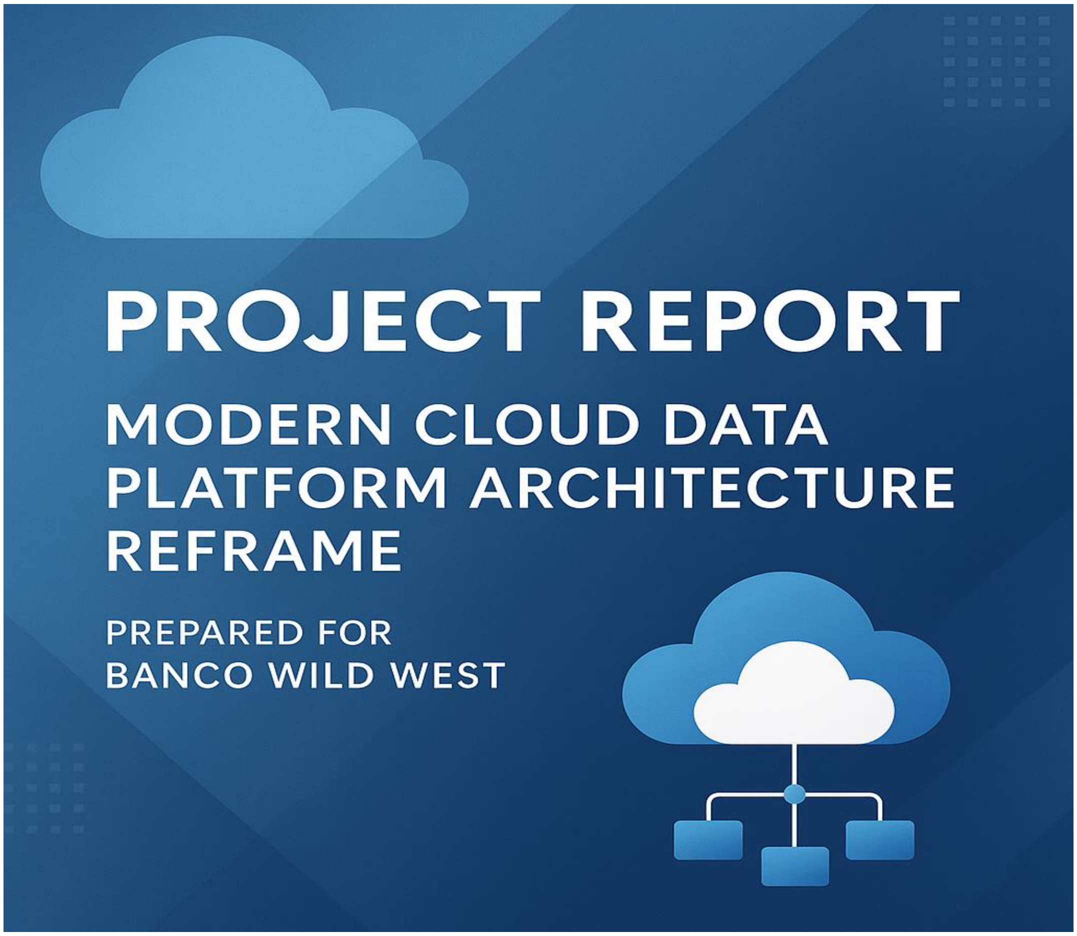

PROJECT REPORT
MODERN CLOUD DATA PLATFORM ARCHITECTURE REFRAME
PREPARED FOR BANCO WILD WEST

ORGANIZATION FOR BUSINESS ANALYTICS PLATFORM

PREPARED BY:

GROUP 2

Shivapreya Durgaprasad, Thriksha Giriraju, Hrushikesh Khandare, Wenting Lu

Srijananie Nagasubburajan, Sahil Salvi, Prasanna Warad

DATE: MAY 8, 2025

INDEX

1. CONTEXT & PROJECT OVERVIEW

Banco Wild West (BWWB), a prominent regional bank, currently relies on outdated IT infrastructure consisting primarily of legacy distributed RDBMS systems, middleware developed in C++/Java, and static data warehousing architectures. These technologies exhibit critical limitations such as poor scalability, frequent outages, limited real-time analytics capabilities, and minimal support for emerging data types. Facing intense competitive pressures from digitally advanced financial institutions, along with evolving regulatory standards and customer demands for personalized, real-time financial services, BWWB has initiated a comprehensive modernization project spearheaded by the newly appointed Chief Information and Data Officer (CIDO). The project's primary objective is to establish a robust, secure, and scalable Azure-based cloud data platform, integrating operational (OLTP) and analytical (OLAP) use cases, while enabling advanced analytics, artificial intelligence (AI), and machine learning (ML) capabilities.

2. KEY PLATFORM REQUIREMENTS

2.1 Technical Requirements:

- Consolidate structured, semi-structured, and unstructured data from diverse sources.
- Support multi-mode data ingestion, including batch, micro-batch, and real-time streaming.
- Establish a Single Source of Truth (SSOT) through comprehensive governance, [1]metadata management, and controlled access.
- Ensure data compliance and security alignment with regulatory frameworks.

2.2 Business Requirements:

- Provide scalable infrastructure that can accommodate future growth and peak usage periods.
- Enable cost-effective operations with optimized resource allocation.
- Facilitate real-time analytics applications including fraud detection and predictive ATM maintenance.
- Deliver hyper-personalized experiences for digital banking customers.
- Support regulatory compliance audits and reporting demands seamlessly.

[1]

3. PLATFORM DECISION: MICROSOFT AZURE

3.1 Evaluation Criteria

Banco Wild West conducted an extensive evaluation of various cloud platforms including AWS, Google Cloud Platform, and Microsoft Azure, assessing factors such as scalability, compliance capabilities, integration with legacy systems, machine learning functionalities, and overall cost-efficiency.

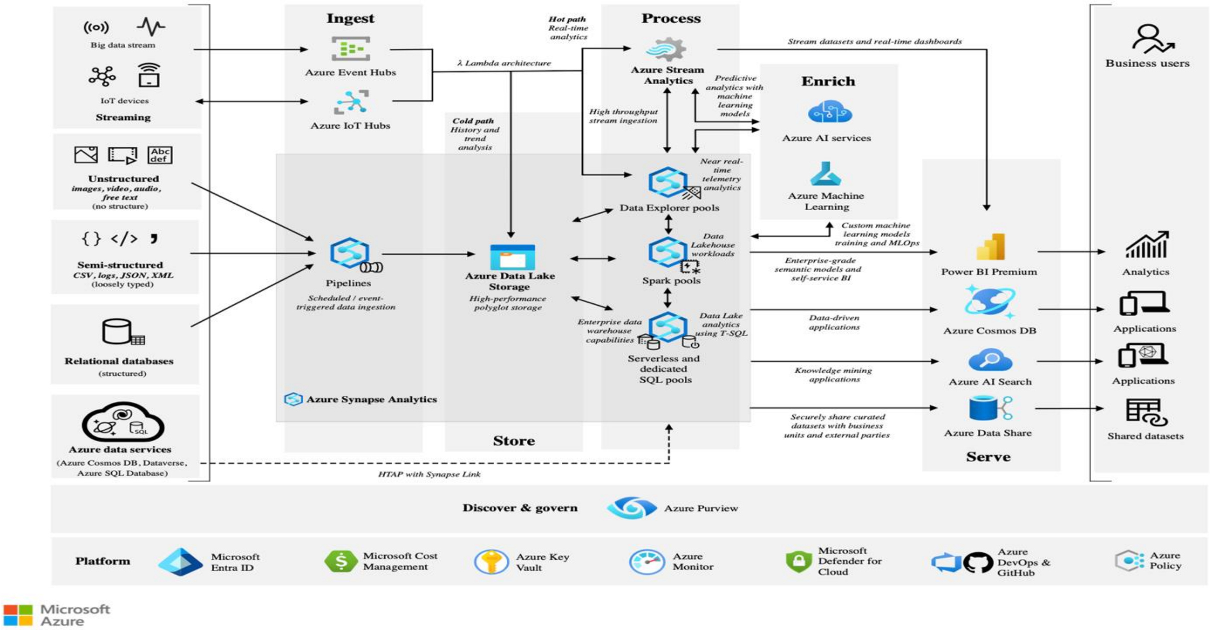
3.2 Advantages of Azure

- **Comprehensive Data Solutions**: Azure provides robust tools such as Azure Synapse Analytics, Azure Databricks, and Azure Data Lake Storage Gen2 (ADLS Gen2) that are well-suited for integrating transactional and analytical workloads within a unified Lakehouse architecture.
- **Security and Compliance**: Azure's built-in compliance tools, including Azure Purview, Azure Defender, Azure Sentinel, and Azure Key Vault, ensure comprehensive data protection, regulatory compliance, and streamlined governance.
- **Scalability and Flexibility**: Azure's dynamic scaling capabilities and flexible pricing models (e.g., pay-as-you-go) align well with the operational and financial objectives of Banco Wild West.
- **Machine Learning Capabilities**: Azure ML and Azure Cognitive Services offer powerful and integrated machine learning environments, enhancing the bank's analytics capabilities.

3.3 Selected Azure Components

- **Azure Synapse Analytics**: Integrated analytics service enabling scalable data warehousing, big data processing, and real-time analytics.
- **Azure Databricks**: Unified analytics platform providing collaborative workspace for data engineering and ML.
- **Azure Data Lake Storage Gen2**: Scalable and secure data storage with hierarchical namespace, optimized for analytics workloads.
- **Azure Machine Learning**: End-to-end ML lifecycle management supporting rapid model development and deployment.
- **Azure Event Hubs and Stream Analytics**: Real-time data ingestion and processing capabilities.

4. DATA PLATFORM ARCHITECTURE (LAKEHOUSE FRAMEWORK)



4.1 Ingestion Layer

- **Batch Ingestion**: Azure Data Factory pipelines integrate data from traditional relational databases like Oracle, SQL Server, and MySQL, enabling scheduled batch ingestion.
- **Real-time Ingestion**: Azure Event Hubs combined with Azure Stream Analytics facilitate continuous data ingestion from ATMs, IoT devices, and social media feeds.
- **API Integration**: Azure Logic Apps and REST API connectors facilitate ingestion from external sources, third-party services, and cloud applications.

4.2 Storage Layer (Bronze)

- **Centralized Storage**: Raw data stored in native formats (CSV, JSON, Avro, XML, PDFs) within Azure Data Lake Storage Gen2, maintaining schema flexibility and supporting diverse data types.
- **Data Versioning**: Managed by Delta Lake to provide version control, audit trails, rollback, and time travel capabilities for enhanced compliance and recovery options.
- **Retention and Compliance**: Implements robust lifecycle management and retention policies ensuring data regulatory compliance.

4.3 Transformation & Processing (Silver)

- **Data Refinement**: Azure Databricks cleanses, standardizes, validates, and enriches data, transforming raw datasets into structured formats.
- **Metadata Management**: Uses Delta Lake for robust metadata tracking, ensuring traceability and governance.
- **Optimized Data Availability**: Prepares data for analytics, ML model training, and operational reporting.

4.4 Curated & Analytics Layer (Gold)

- **High-Performance Analytics**: Azure Synapse SQL pools store optimized, structured, and aggregated data tables designed for fast analytics and real-time querying.
- **Analytical Insights**: Powers business intelligence, predictive analytics, and real-time dashboarding.
- **Machine Learning Integration**: Enables seamless extraction of prepared data for ML models used in fraud detection, credit scoring, and personalized recommendations.

4.5 Consumption Layer

- **Interactive Dashboards**: Power BI Premium offers enterprise-wide visualization tools for executives and operational analysts.
- **Secure APIs**: Azure API Management provides secure, reliable access to backend data services for internal applications and customer-facing solutions.
- **Real-time Data Access**: Azure Cosmos DB delivers ultra-low latency, real-time data queries to support dynamic, responsive customer experiences.

4.6 Governance & Security

- **Data Cataloging and Lineage**: Azure Purview centralizes metadata management, data lineage tracking, and sensitive data classification.
- **Credential Management**: Azure Key Vault securely manages sensitive information including encryption keys, certificates, and credentials.
- **Access Control**: Role-based access control (RBAC) enforced through Azure Active Directory, ensuring precise, role-appropriate data access.
- **Threat Detection and Compliance Monitoring**: Azure Defender and Azure Sentinel provide continuous monitoring, threat detection, and real-time alerts to proactively manage compliance and security risks.

5. DATA MIGRATION STRATEGY

5.1 Data Assessment and Inventory

- Conduct comprehensive audits to identify, catalog, and classify existing datasets according to data type (structured, semi-structured, unstructured), sensitivity, and regulatory compliance needs.
- Develop detailed data inventory documentation specifying sources, formats, volume estimates, dependencies, and usage patterns to facilitate informed migration planning.

5.2 Data Extraction and Initial Migration

- Utilize Azure Data Factory for initial bulk data transfer, setting up automated ETL (Extract, Transform, Load) pipelines from legacy databases like Oracle, SQL Server, MySQL, and Teradata to Azure Data Lake Storage Gen2.
- Establish and validate connections, configurations, and performance metrics, ensuring data integrity through checksum and row count verification.

5.3 Real-Time Data Integration

- Activate real-time ingestion through Azure Event Hubs, handling data streams from ATMs, sensor telemetry, customer interactions, and transaction logs.
- Integrate Azure Stream Analytics to process streaming data in real-time, ensuring minimal latency and continuous data availability.

5.4 Data Validation and Transformation

- Deploy Azure Databricks notebooks and jobs to perform thorough data cleansing, schema alignment, deduplication, and enrichment processes, transforming data into structured, analytics-ready formats within the silver zone.
- Execute rigorous validation protocols, including reconciliation checks, data profiling, and quality assessments, to ensure data consistency and accuracy.

5.5 Parallel Operation and Testing

- Operate legacy and new systems concurrently during a transition period to guarantee business continuity, conducting side-by-side validation to identify discrepancies and rectify potential data quality or processing issues.
- Utilize Delta Lake versioning and rollback capabilities for risk mitigation, allowing immediate reversions if inconsistencies or errors arise.

5.6 Final Migration and Cutover

- Gradually reduce reliance on legacy systems post successful validation, systematically transitioning operational processes, reporting, and analytics workloads entirely to the new Azure-based platform.
- Schedule cutover during off-peak periods to minimize business disruptions, employing automated monitoring and alerting tools for post-migration observation and rapid response capabilities.

5.7 System Decommissioning

- Following stable and successful operation of the new platform, initiate controlled decommissioning procedures for legacy systems.
- Implement archival strategies for historical and infrequently accessed data, ensuring regulatory compliance and cost-efficient long-term storage.
- Document and communicate decommissioning activities transparently to all stakeholders, securing necessary approvals and maintaining detailed records for audit and compliance purposes.

6. ML USE CASES TO SUPPORT BUSINESS GOALS

6.1 Fraud Detection

- **Objective:** Enhance real-time transaction monitoring capabilities to quickly identify and prevent fraudulent activities.
- **Techniques:**
    - Anomaly detection algorithms for identifying unusual transaction patterns.
    - Supervised classification models trained on historical labeled data to recognize fraudulent transactions.
- **Implementation:** Utilize Azure Databricks for real-time feature extraction and Azure ML for model training and deployment. Integrate with Azure Event Hubs for immediate transaction processing.

- **Outcome:** Reduced financial losses, improved customer trust, and compliance with regulatory requirements.

6.2 Credit Risk Scoring

- **Objective:** Streamline and improve accuracy in loan application processing by predicting default risks.
- **Techniques:**
  - Machine learning models leveraging customer demographics, credit histories, employment details, and transaction patterns.
  - Logistic regression, decision trees, and gradient boosting algorithms to enhance predictive performance.
- **Implementation:** Develop predictive models using Azure ML and Azure Databricks, integrated directly with operational loan approval workflows.
- **Outcome:** Increased loan processing efficiency minimized default rates, and enhanced decision-making accuracy.

6.3 Personalized Product Recommendations

- **Objective:** Deliver personalized banking experience by offering tailored financial products and services.
- **Techniques:**
  - Collaborative filtering and content-based recommendation algorithms.
  - Customer segmentation and predictive analytics to anticipate financial needs.
- **Implementation:** Utilize Azure Databricks for data preparation and Azure ML for developing and continuously updating recommendation models.
- **Outcome:** Improved customer engagement increased cross-selling opportunities, and enhanced customer retention.

6.4 Predictive ATM Maintenance

- **Objective:** Minimize ATM downtime by predicting and proactively addressing maintenance needs.
- **Techniques:**
  - Regression models using ATM telemetry and operational data.
  - Time-series forecasting to predict equipment failures or performance degradation.
- **Implementation:** Leverage Azure IoT Hub for telemetry data ingestion, Azure Databricks for feature engineering, and Azure ML for predictive analytics.

- **Outcome:** Reduced operational disruptions, lowered maintenance costs, and enhanced customer service availability.

6.5 Automated Customer Support

- **Objective:** Enhance customer service efficiency through AI-driven chatbot solutions.
- **Techniques:**
  - Natural language processing (NLP) using Azure Cognitive Services.
  - Machine learning algorithms to understand and respond effectively to customer inquiries.
- **Implementation:** Develop and deploy intelligent chatbot interfaces through Azure Bot Service integrated with Azure Cognitive Services.
- **Outcome:** Improved customer support responsiveness, reduced service center workload, and increased customer satisfaction.

6.6 Anti-Money Laundering (AML)

- **Objective:** Strengthen regulatory compliance by proactively identifying potential money laundering activities.
- **Techniques:**
  - Graph analytics to detect suspicious transaction networks.
  - Anomaly detection and clustering algorithms to identify unusual financial behavior.
- **Implementation:** Employ Azure Databricks and Azure Graph APIs for comprehensive analysis and visualization of transaction networks.
- **Outcome:** Enhanced regulatory compliance minimized financial and reputational risks, and improved detection efficiency.

7. COMPLIANCE & GOVERNANCE

7.1 Regulatory Alignment

- Comprehensive adherence to regulatory frameworks including IRS, GDPR, HIPAA, and NIST 800-53.
- Regular compliance assessments, gap analysis, and continuous improvement protocols.

7.2 Data Lifecycle Management

- Implement strict lifecycle management policies within Azure Data Lake Storage, including data retention, archiving, and deletion policies compliant with regulatory standards.
- Automate lifecycle transitions based on usage patterns and compliance requirements.

7.3 Access Control and Identity Management

- Integrate role-based access control (RBAC) through Azure Active Directory to manage and audit user permissions systematically.
- Enforce multi-factor authentication (MFA) and conditional access policies for sensitive data access.

7.4 Data Protection and Encryption

- Ensure data encryption at rest and in transit using Azure encryption services and Azure Key Vault.
- Support customer-managed keys and Bring Your Own Key (BYOK) options for sensitive data encryption.

7.5 Monitoring and Threat Detection

- Deploy Azure Defender and Azure Sentinel for continuous, real-time monitoring of system activity and potential security threats.
- Implement automated alerting mechanisms and incident response procedures to quickly address detected threats and anomalies.

7.6 Auditability and Reporting

- Use Azure Monitor and Azure Log Analytics for centralized logging and audit trails, facilitating transparency and regulatory audits.
- Automated compliance reporting processes with Azure services to provide timely and accurate documentation for regulatory review.

7.7 Data Governance and Metadata Management

- Utilize Azure Purview for comprehensive data governance, metadata cataloging, data lineage tracking, and sensitive data classification.

- Maintain clearly defined data stewardship roles and responsibilities to ensure adherence to governance policies and standards.

8. ROADMAP FOR EXECUTION

Quarter	Key Milestones	Detailed Activities
Q1	Infrastructure setup, Data Inventory, Access Control, ADLS provisioning	- Setup foundational Azure cloud infrastructure (ADLS, Synapse, Databricks). - Establish security protocols with Azure AD, Key Vault. - Conduct comprehensive inventory and categorization of existing data assets. - Configure data access controls and permissions.
Q2	Batch ingestion pipelines, Bronze/Silver setup, Data Governance via Purview	- Develop and deploy Azure Data Factory pipelines for batch data ingestion. - Set up Bronze (raw) and Silver (cleaned) data layers in ADLS. - Implement initial data governance framework using Azure Purview for metadata management and lineage tracking. - Validate data ingestion quality and performance.
Q3	Streaming pipelines, ML Model Prototypes, Synapse Integration	- Activate streaming data pipelines using Azure Event Hubs and Stream Analytics. - Develop machine learning prototypes for fraud detection, credit scoring, and ATM maintenance using Azure ML. - Integrate Azure Synapse for seamless data analytics and reporting. - Test and optimize real-time processing performance.
Q4	Final BI Dashboards, API Endpoints, Compliance Automation, Monitoring Setup	- Launch final Power BI dashboards for executive and operational decision-making. - Expose secure API endpoints via Azure API Management for internal/external application integration. - Automate regulatory compliance reporting and auditing tasks. - Establish comprehensive monitoring and alerting infrastructure with Azure Monitor and Sentinel.

9. CHALLENGES AND MITIGATION

Challenge	Mitigation Strategy	Detailed Explanation
Legacy Data Complexity	Utilize Azure Data Factory + Delta Lake versioning for smooth migration	Employ modular migration approach, leveraging ADF pipelines and Delta Lake versioning to manage schema changes, ensure data integrity, and facilitate rollback capabilities.
Real-Time Latency	Implement Azure Event Hubs + Synapse Link + Cosmos DB integration	Deploy Azure Event Hubs for event streaming, utilize Azure Synapse Link for real-time analytics, and synchronize data with Azure Cosmos DB for low-latency queries and fast data retrieval.
Metadata Overload	Deploy Azure Purview with manual curation tags	Implement Azure Purview as a central metadata management tool, establish manual and automated tagging to simplify metadata management, enhance discoverability, and ensure governance.
Governance Setup	RBAC, Tag-based policies in Purview, ADLS ACLs	Establish role-based access controls integrated with Azure AD, use tag-based policies within Azure Purview for granular access control, and apply ACLs in Azure Data Lake Storage to secure data assets effectively.
Cost Control	Archive tier for cold data, Synapse serverless pools	Implement lifecycle management policies to move infrequently accessed data to Azure's low-cost archive tiers, leverage Synapse serverless SQL pools for cost-effective ad-hoc queries and optimize resource usage through auto-scaling policies.

10. COST ANALYSIS

Component	Azure Service	Monthly Cost Estimate	Detailed Justification
Storage	ADLS Gen2 (500TB)	$10,000	Supports vast data storage requirements; leverages lifecycle management to optimize storage tiers, ensuring cost-efficiency.
Compute	Databricks + Synapse	$7,500	Provides scalable data processing and analytics capacity; optimized using reserved instances, auto scaling, and shared pools.
Ingestion	ADF + Event Hubs	$2,500	Efficiently manages batch and real-time data ingestion with optimized scheduling and resource allocation.
Visualization	Power BI Premium (1000 users)	$3,500	Offers comprehensive BI capabilities for broad user base; managed through role-specific dashboards and user segmentation.
ML/AI	Azure ML + Cognitive Services	$5,000	Supports advanced ML and AI functionalities required for predictive analytics, fraud detection, and personalization; cost controlled via spot instances and efficient lifecycle management.
Governance	Purview + Defender + Sentinel	$1,200	Ensures robust governance and proactive threat detection; utilizes centralized logging, shared resources, and automated alerts for cost-effective management.
Security	Key Vault + RBAC + Monitor	$800	Secures sensitive information and credentials; integrated RBAC and continuous monitoring tools provide essential protection with cost-effective resource sharing.
Total		**~$30,500/month**	Comprehensive budget aligns strategically with project objectives, ensuring scalability, compliance, and optimal resource usage.

10.1 Cost Optimization Strategies:

- Regular review of Azure Cost Management dashboards.
- Implementation of automated alert systems for spending thresholds.
- Periodic reassessment of service usage to identify savings opportunities.
- Lifecycle and archiving policies for older data, reducing unnecessary storage and compute overhead.

10.2 Projected Return on Investment (ROI):

- Significant operational cost reduction compared to legacy infrastructure (estimated 35-40%).
- Enhanced scalability and reduced downtime leading to increased customer satisfaction and business growth.
- Improved compliance posture and reduced regulatory risk exposure, contributing to overall financial stability.

11. CONCLUSION

Banco Wild West's strategic initiative to modernize its data platform using Microsoft Azure positions the bank to meet contemporary market demands, improve operational efficiency, and enhance regulatory compliance capabilities. The comprehensive Azure-based Lakehouse architecture proposed in this project facilitates integrated real-time analytics, machine learning applications, and robust data governance. With a structured implementation roadmap, detailed risk mitigation strategies, and optimized cost management practices, Banco Wild West is set to achieve substantial operational savings, heightened customer satisfaction, and a competitive edge in the financial services landscape.

12. REFERENCES

- Azure. (n.d.). Microsoft Azure. Retrieved from https://azure.microsoft.com/
- Microsoft. (n.d.). Azure Architecture Center. Retrieved from https://learn.microsoft.com/en-us/azure/architecture/
- Microsoft Tech Community. (n.d.). Retrieved from https://techcommunity.microsoft.com/
- Turck, M. (2021). Data & AI Landscape 2021. Retrieved from https://mattturck.com/data2021/
- Microsoft Azure. (n.d.). Cloud Compliance. Retrieved from https://azure.microsoft.com/en-us/resources/cloud-compliance/

- High Scalability. (n.d.). Retrieved from https://highscalability.com/