**Section 1: Core Architecture Concepts**

**Q1. What is your approach to designing a modern data architecture?**  
**A:** I follow a layered architecture:

* **Data Ingestion Layer** (batch/stream using ADF, Event Hubs)
* **Data Storage Layer** (data lake using OneLake/Data Lake Gen2)
* **Processing Layer** (Synapse, Databricks, Spark)
* **Semantic Layer** (Power BI, Analysis Services)
* **Consumption Layer** (dashboards, APIs)  
  I consider scalability, data governance, cost, latency, and security as core design principles.

**Q2. How do you choose between a data warehouse and a data lake?**  
**A:**

* **Data Lake** is ideal for raw, semi-structured/unstructured data, big data analytics, and machine learning.
* **Data Warehouse** (like Synapse or Snowflake) is best for structured, curated data used in BI and reporting.  
  I often use a **lakehouse** architecture to get benefits of both.

**Q3. What factors do you consider for database platform selection (e.g., Azure SQL vs Cosmos DB)?**  
**A:**  
I evaluate:

* Data structure and access patterns
* Transaction vs analytical workloads
* Consistency vs availability needs
* Cost, scalability, latency  
  Example: I’d use Cosmos DB for globally distributed NoSQL workloads with low-latency reads and writes, and Azure SQL MI for transactional apps needing strong consistency and T-SQL support.

**🔸 Section 2: Cloud & Azure Architecture**

**Q4. How do you implement data governance in Azure?**  
**A:**  
Using **Microsoft Purview** for:

* Data discovery and classification
* Lineage and impact analysis
* Business glossary  
  Along with **RBAC**, **Azure Policy**, **Private Endpoints**, **Managed Identities**, and **sensitivity labels** to enforce governance.

**Q5. How would you design a data platform for real-time analytics?**  
**A:**

* **Ingestion:** Azure Event Hubs / IoT Hub
* **Stream processing:** Azure Stream Analytics or Apache Flink
* **Storage:** Delta tables in Lakehouse
* **Real-time dashboards:** Power BI with Direct Lake or Kusto (ADX)  
  Also ensure **horizontal scalability**, **message replay**, and **event ordering** are supported.

**Q6. How do you ensure security across a distributed data architecture?**  
**A:**

* **Data encryption:** At rest (Azure-managed keys/CMK) & in transit (TLS)
* **Access control:** RBAC, ACLs, Managed Identity
* **Networking:** VNET integration, Private Link, NSGs
* **Auditing & monitoring:** Azure Monitor, Log Analytics, Defender for Cloud

**🔸 Section 3: Data Modeling & Integration**

**Q7. Describe a scenario where you optimized a data model for analytics.**  
**A:**  
At my previous role, we redesigned a star schema to avoid snowflaking and denormalized frequent joins. We also optimized date/time columns, reduced cardinality, and implemented aggregations. This improved Power BI report load time from 12s to under 3s.

**Q8. What are the best practices for managing schema evolution in a data lake?**  
**A:**

* Use Delta Lake with schema enforcement
* Track schema versions using Git or Purview
* Perform validation before merge
* Isolate breaking changes with branching or versioned folders

**Q9. How do you handle slowly changing dimensions (SCD) in a cloud ETL pipeline?**  
**A:**

* **Type 1:** Overwrite
* **Type 2:** Insert new row with versioning (using ADF or Synapse pipeline)
* **Type 3:** Track limited history in the same row  
  In Delta Lake, use MERGE INTO with conditions to manage SCD logic efficiently.

**🔸 Section 4: Big Data & Advanced Scenarios**

**Q10. How do you handle data duplication and quality in large-scale systems?**  
**A:**

* **Deduplication:** Use hashes or surrogate keys, Delta Lake MERGE or DISTINCT
* **Data Quality:** Use Great Expectations or custom validation pipelines
* **Monitoring:** Track data freshness, row counts, null rates, schema drifts

**Q11. What’s your approach to building a multi-tenant data platform?**  
**A:**

* Use workspace or schema separation
* Metadata-driven pipelines
* Enforce RBAC and row-level security
* Consider cost segregation (billing tags or isolated storage)

**Q12. How would you migrate Redshift to Synapse?**  
**A:**

* Export Redshift to S3 → Move to Azure Blob via AzCopy
* Convert schema using AWS SCT or manual scripts
* Use ADF or COPY INTO in Synapse
* Validate row counts and performance
* Rebuild pipelines and BI integrations

**🔸 Section 5: Strategic & Leadership**

**Q13. How do you align data architecture with business strategy?**  
**A:**  
By engaging stakeholders early to understand KPIs, reporting needs, and compliance goals. I map these to architecture decisions (e.g., real-time data for operations, historical data for forecasting) and ensure agility through modular, scalable design.

**Q14. How do you evaluate new data technologies (e.g., Microsoft Fabric)?**  
**A:**  
I assess based on:

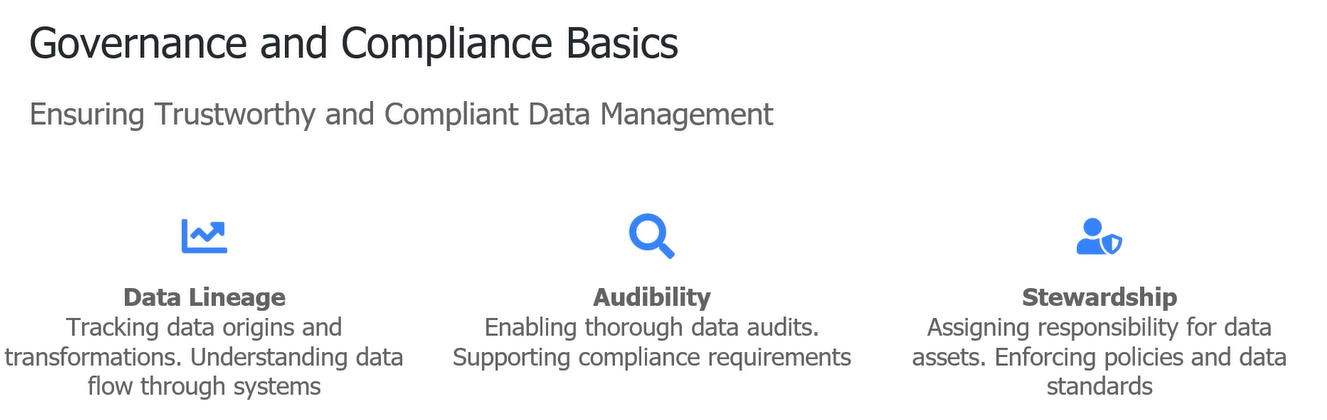
* Fit for existing use cases
* Cost vs benefit
* Integration with current stack
* Team skill readiness  
  I often run **PoCs** with measurable metrics like performance, governance support, and usability.

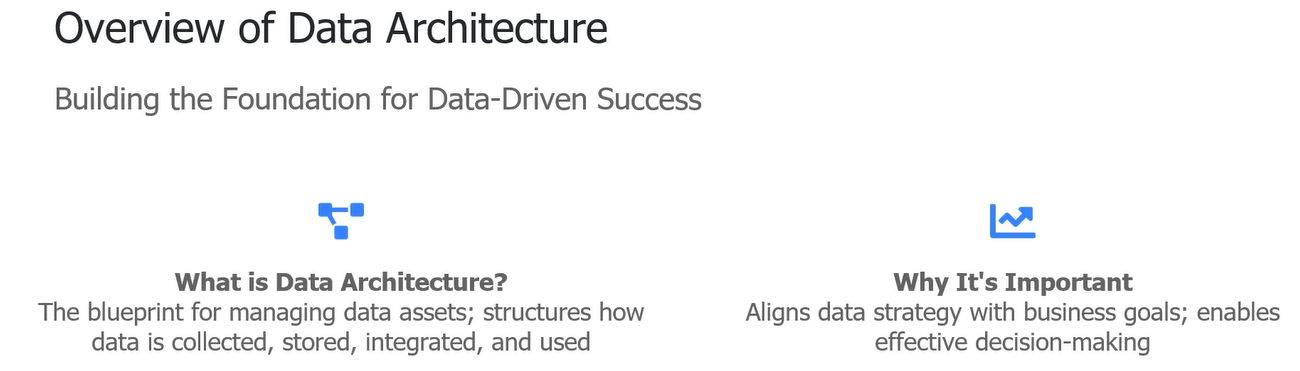
**Q15. Describe a challenge you faced as a data architect and how you resolved it.**  
**A:**  
In a recent project, data silos across departments caused inconsistent reporting. I proposed a unified data platform using a lakehouse model, led data catalog adoption (Purview), and standardized models. Result: data reusability increased and time-to-insight decreased by 40%.



A close up of a plant

AI-generated content may be incorrect.



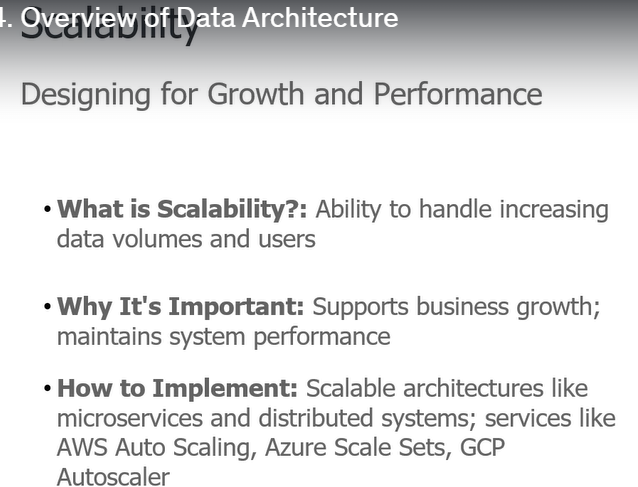


A close-up of a data architecture

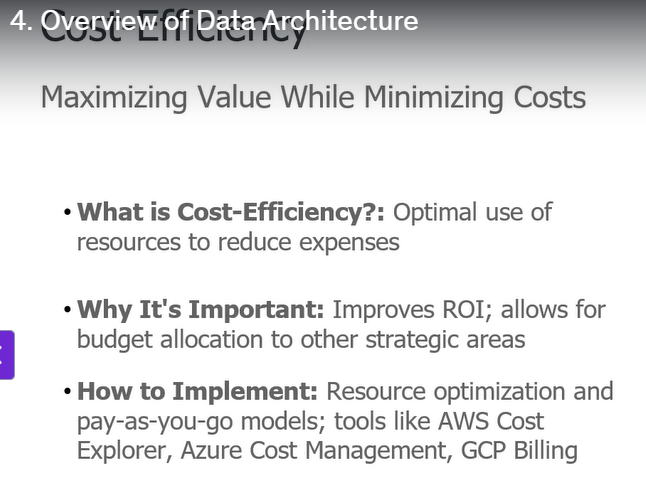
AI-generated content may be incorrect.

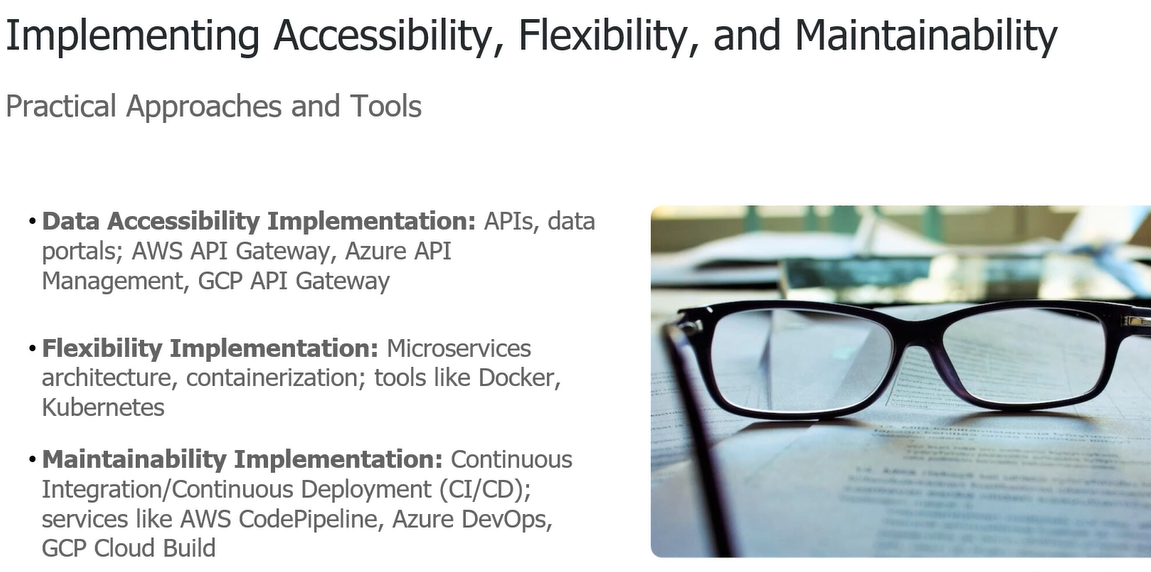
A screenshot of a computer

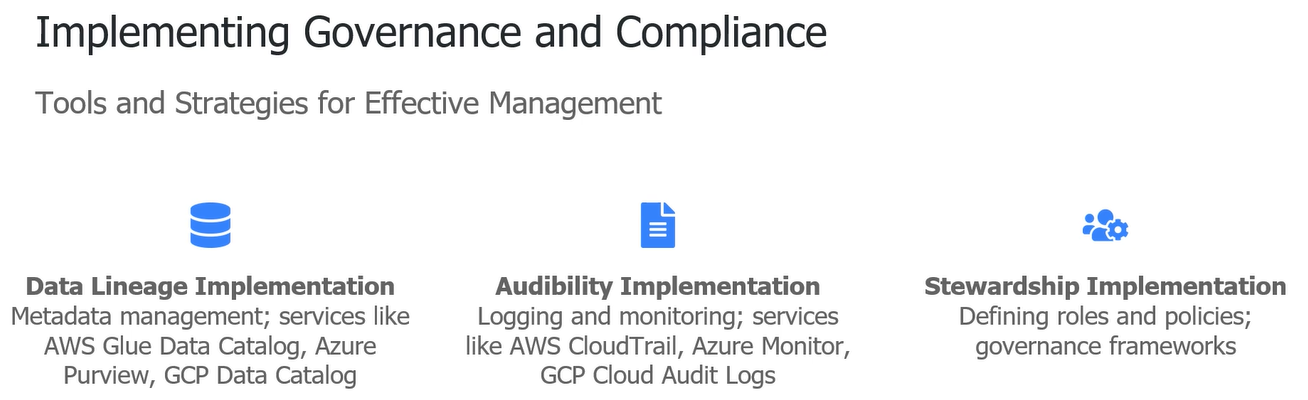
AI-generated content may be incorrect.

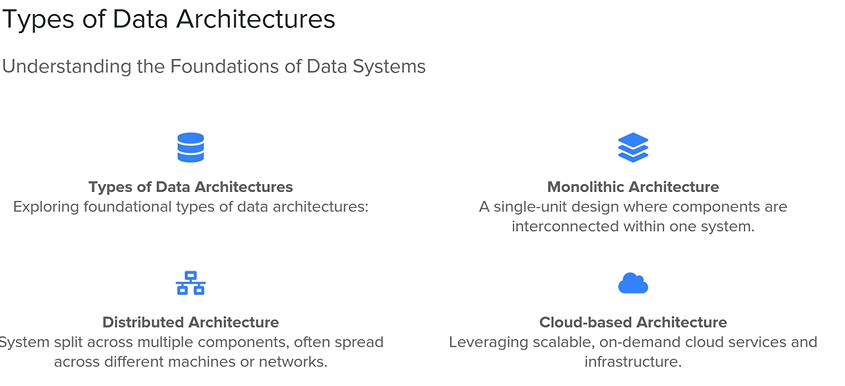






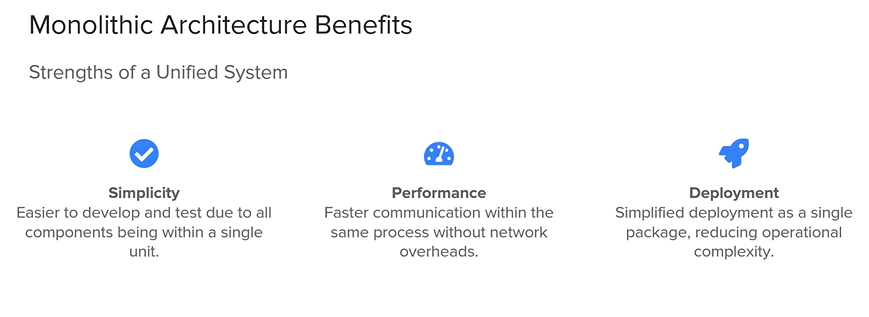






A screenshot of a computer

AI-generated content may be incorrect.



A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a device

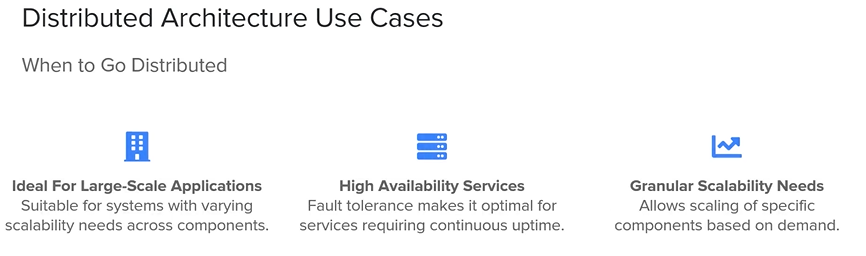
AI-generated content may be incorrect.

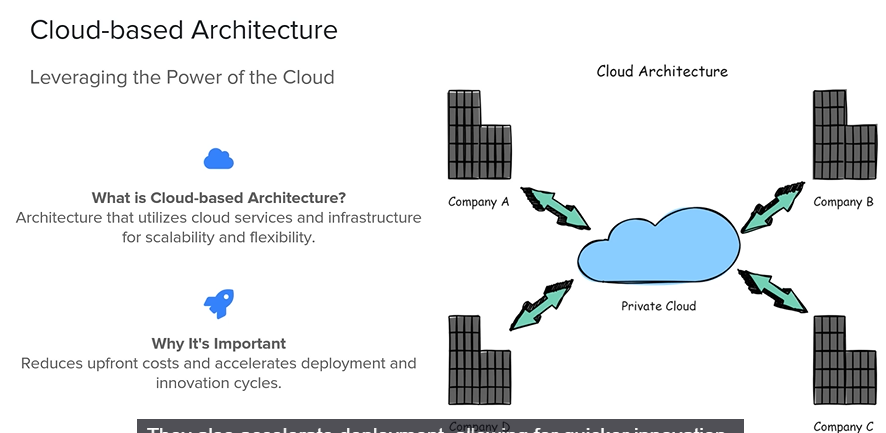
A diagram of a computer network

AI-generated content may be incorrect.

A screen shot of a computer

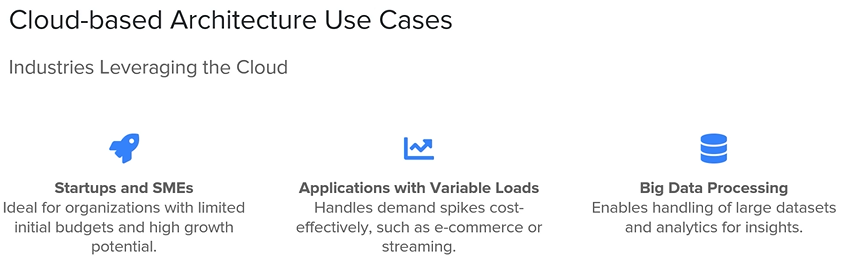
AI-generated content may be incorrect.

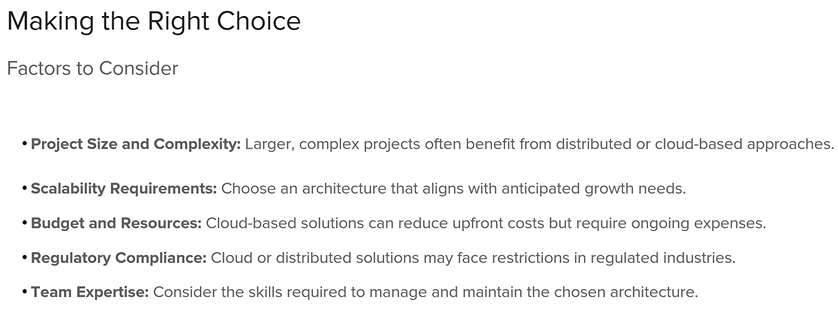


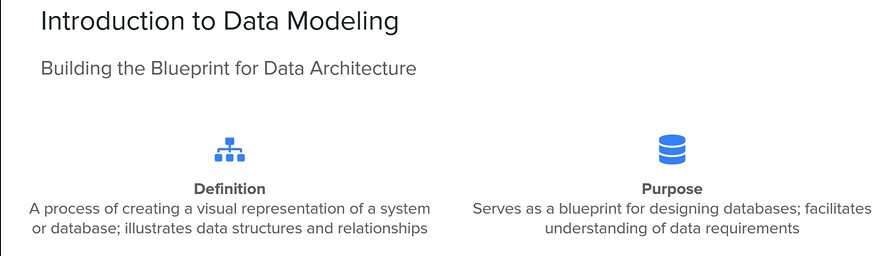


A screen shot of a computer

AI-generated content may be incorrect.







A screenshot of a data modeling

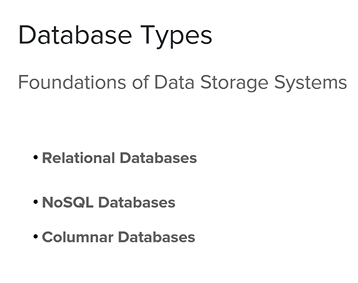
AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

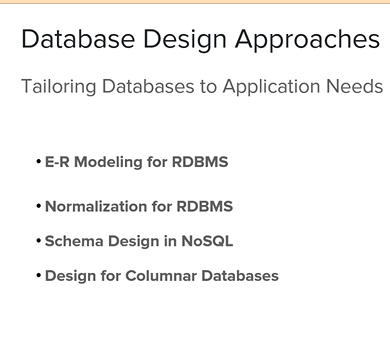
A diagram of a model

AI-generated content may be incorrect.



A close-up of a white background

AI-generated content may be incorrect.

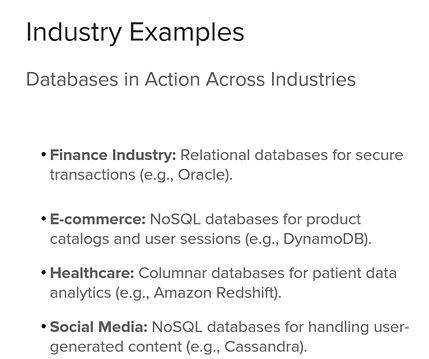


A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

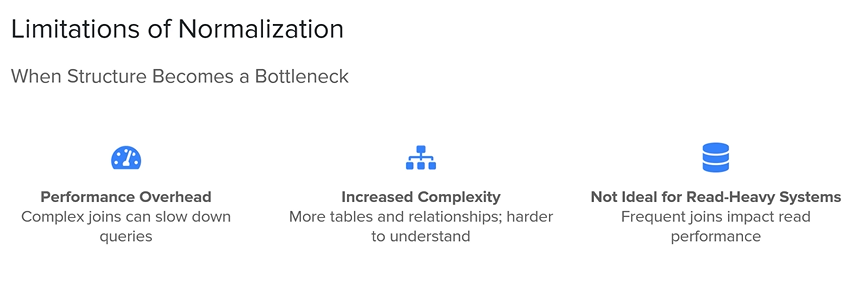


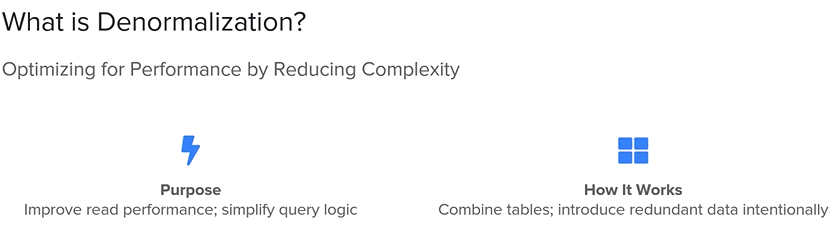
A white background with black text

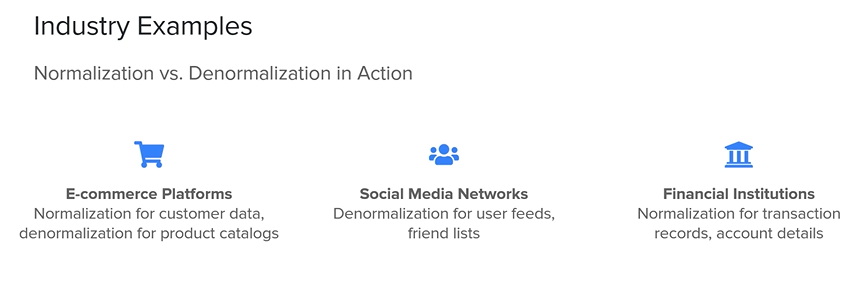
AI-generated content may be incorrect.

A logo of a company

AI-generated content may be incorrect.







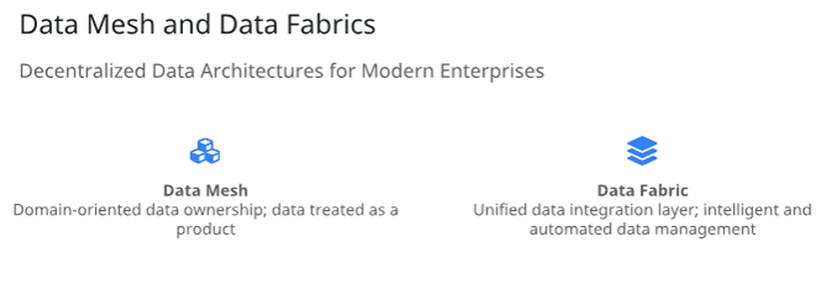
ELT VS ETL

BATCH Data processing

Real time processing

Data lakes vs data warehouse

Data mesh and data fabrics



A close-up of a business plan

AI-generated content may be incorrect.

