Data Platform Architecture & Strategy considerations for Business Analytics & Reporting

The Analytics and data strategy in an enterprise depend on many factors. Almost all the times the decision to adopt a particular strategy depend on many factors both internal and external to the organization.

Below is the list of few considerations that are important in designing a data and analytics platform

- ❑ Types of Data Architecture – Data Federation vs Replication (ETL)
- ❑ Landscape – Modernization vs upgrade
- ❑ Storage – Public Cloud vs Private Cloud, On-Premise
- ❑ Integration – Source Systems, Meta-data Management, Reporting Tools, Governance, other Services
- ❑ Data Categorization
- ❑ Other considerations – Cost Considerations, Business and Data Teams, Legacy systems

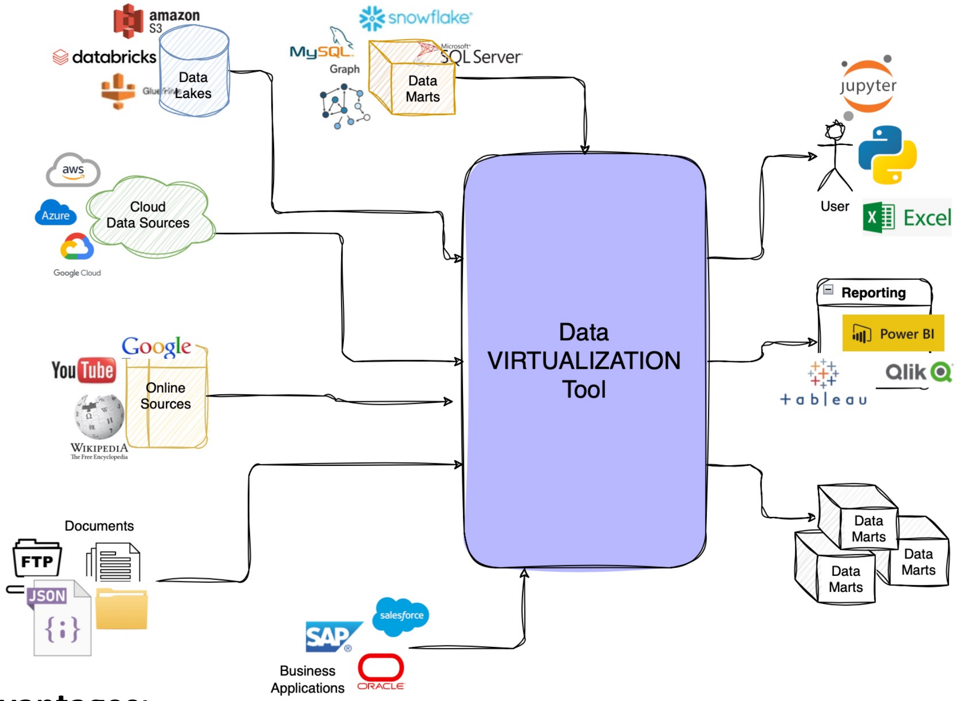
Author: Ravi Akasapu

Types of Data Architecture –
Data Federation vs Replication (ETL)

Author: Ravi Akasapu

Data Virtualization

Data Virtualization is where multiple sources of data is accessed virtually, without the need for actual movement. The data is physically present in disparate data sources but can be accessed & integrated for presentation.



Advantages:

- Use of data virtualization resulting in
  - Reduced storage
  - Easy data modelling while combining multiple sources of data
- Easy and Flexible data access from multiple platforms and formats
- Reduces data redundancy and improved agility
- Real-time and accurate data
- Improved data security
- Lower cost of Infrastructure and storage

Disadvantages:

- Required uniform data formats for easy access
- Restricted source system access can raise issues in data access
- Complex/incomplete/inconsistent source data can lead to inefficient virtual models
- Reporting on large volume of data can introduce time delays

Author: Ravi Akasapu

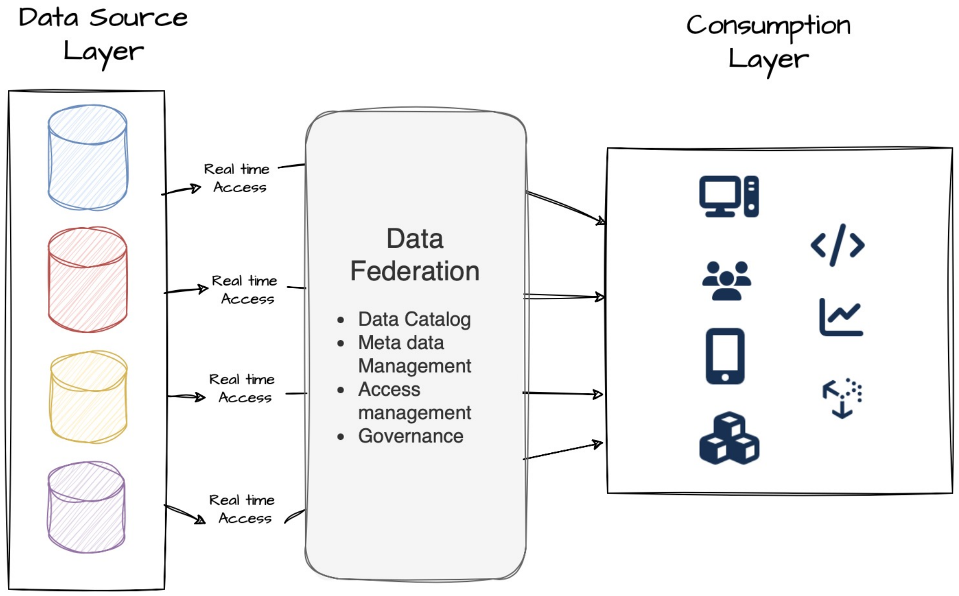
Data Federation

Data Federation is a data Management strategy that queries data directly from various sources to be used in reporting or in other processes. It is similar to data virtualization but instead of accessing the virtually from various sources, federation will provide access from one common data model.

There are several types of Data Federation architectures available: Full Mesh Federation, Hybrid Mesh Federation and Governed Mesh Federation

There are additional advantages like:

- Improved Metadata Data Management and Governance
- Improved access management and security
- Ability to run query pushdown and parallelization
- Complex transformations in Queries at source level
- etc..



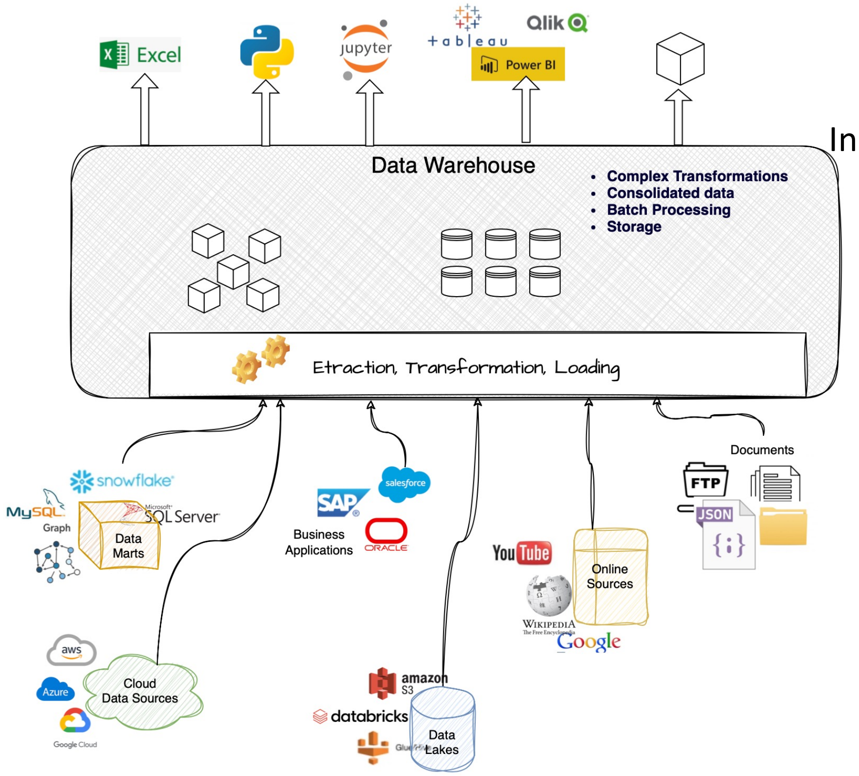
Data Federation

- Data Catalog
- Meta data Management
- Access management
- Governance

Author: Ravi Akasapu

Data Replication(ETL/ELT)/Data Warehouse

Instead of accessing the data virtually a **Data Warehouse** replicates the data from various sources and stores or consolidates in a central location.



A Data warehouse uses various connectors and data management tools to bring the data into the warehouse. It offers features like

- Connection management
- ETL/ELT based data replication. Offers both scheduled/real time
- Consolidated tool Analytics/Reporting or further streaming of the data

Advantages:

- One source of truth for all data and reporting
- complex transformations
- data cleaning techniques can be applied for improper or incorrect data
- Unified access management

Disadvantages:

- Data Duplication and redundancy
- Increased storage and costs
- Might become Complex modelling
- Adds to more infrastructure costs and

Author: Ravi Akasapu

Which approach is best ???

- No one approach is best for all scenarios. It is always depends on case-to-case basis and a Hybrid approach will serve most of the use cases.
- Below is list of use cases and advantages of each strategy.

Use Case	Data Virtualization	Data Warehousing
Real-time Data Access	Suitable for real-time data access.	May not be ideal due to ETL process latency.
Historical Data Analysis	Less effective, and time taking to process huge volume of data.	Ideal for historical data analysis. Can process huge volume data using aggregates
Data Integration from Multiple Sources	Excellent for integrating data from multiple sources without moving data.	Requires ETL processes to consolidate data
Rapid Prototyping & Testing	Perfect for quick setup and testing.	Slow due to model setup and batch processing
Complex Queries & Reporting	Not suitable for complex query processing	Best suitable as it is possible to create complex transformations
Data Consistency & Quality Control	Depends on underlying data sources	Data quality can be maintained within the warehouse
Cost Efficiency	No/minimal Storage costs. Processing costs	Can be expensive due to storage and maintenance.
Data Security & Governance	Hard to manage across different sources.	Easier to secure and govern centrally.
Scalability	Scales as needed but struggles with processing	Scales well with more data and users.
Agility in Changing Data Sources	Highly agile	Less agile. Need effort in ETL
Long-Term Data Retention	Not possible. Cannot retain large data	Ideal for long term retention. Can archive data as needed
Regulatory Compliance	Complex. Mostly depends on source systems	Easy and centralized.

Author: Ravi Akasapu

Landscape

Modernization vs Upgrade

Author: Ravi Akasapu

Landscape Approaches

Every organization that has a data analytics platform uses products & tools from several vendors within the landscape. The data movement happens between and across multiple products either real-time or batch processing.

The design of the landscape enables the enterprise to process, analyze and derive insights from the data.

There are 2 primary approaches to while taking decisions with regard to data analytics platforms

- **Modernization -** Complete overhaul and transformation of the existing data architecture.
- **Upgrade** - Incremental improvements to the current landscape utilizing current investments in the architecture.

Author: Ravi Akasapu

Modernization

Complete overhaul and transformation of the existing data architecture. This requires organization to bring latest and completely different technologies into the landscape.

- Need for advanced analytics capabilities like real-time, agile reporting
- Move to cloud technologies and modern tools
- Integration with AI & ML and other innovations
- Scalability and Flexibility
- Future proofing the data architecture
- Outdated infrastructure. non-availability or limited technology support
- Employing proper Governance tools for Data and Access

Challenges

- Higher initial cost
- Significant change management (Development, Training, Adoption etc..)
- Legacy systems

Upgrade

Continue to use existing tools and employing improvements without replacing.

- Significant investments in architecture (Internalization, developer expertise, user adoption)
- Minimize disruptions to the business operations
- Cost constraints
- Complicated legacy systems affecting future integrations

Challenges

- Technical debt accumulation
- Unable to meet changing business and technological advancements

Author: Ravi Akasapu

Modernization vs Upgrade

- Again, it all depends. A Hybrid approach suits most architectures and scenarios
- It is always better to add New tools with latest technologies than updating existing tools

Use Case	Modernization	Upgrade
Data	Complicated to transfer data. Can be challenging due to more components during the changes.	Since the data is in the same place, easy to manage the upgrade
Technology Stack	Adopting a new and modern technologies and platforms add more value	Adds incremental features/updates.
Business Alignment	Aligns the data architectures to current & future business objectives	Caters to current business processes, but can be fitted to align. Need more efforts
Governance, Compliance and Security	Can improve the Governance, Compliance and security significantly by adding more and appropriate tools. Follows the changing regulatory needs.	Adds value by adopting existing processes to new regulatory changes. May not be as efficient and effective.
Flexibility and experience	Adds more flexibility and improves user experience and engagement significantly	Could be improved and adds value

Author: Ravi Akasapu

Storage

Public Cloud , Private Cloud, On-Premise

Author: Ravi Akasapu

Storage

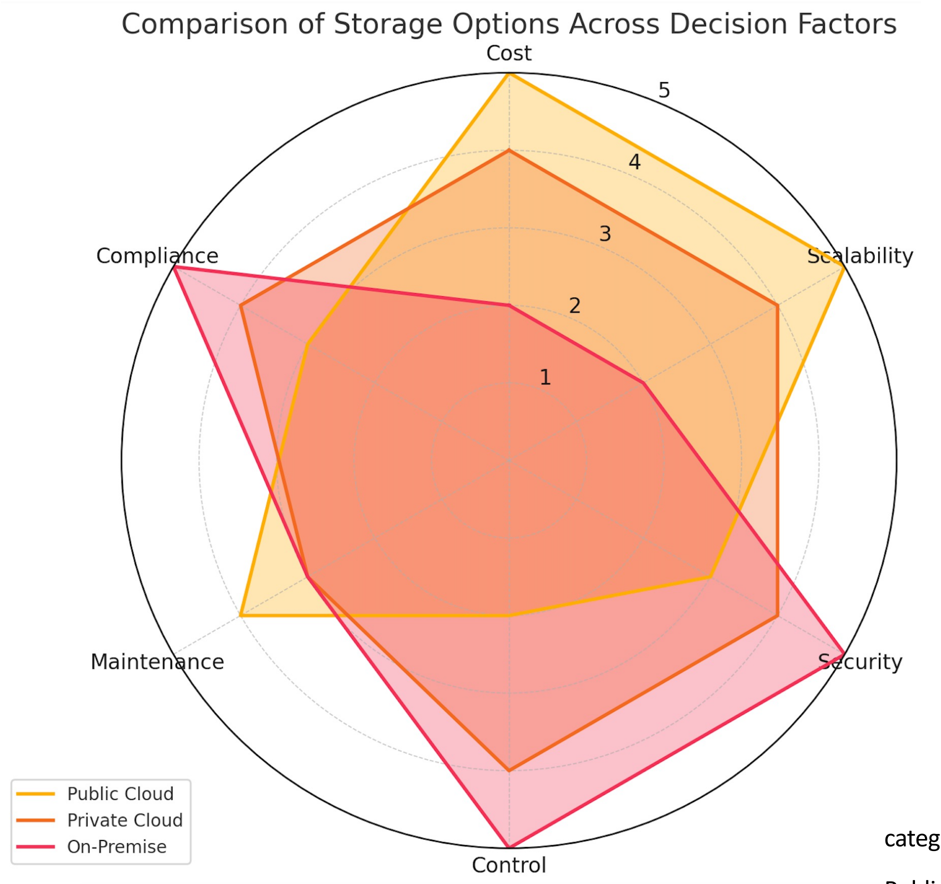
Storage is a critical component of an enterprise's data analytics strategy. Choosing the right storage option impacts scalability, security, cost, and performance. Storage also affects the architecture and connectivity of tools within the landscape.

The 3 storage options that are available : **Public Cloud**, **Private Cloud**, and **On-Premise**

	Public Cloud	Private Cloud	On-Premise
Provider	Provided by third party vendors on a share infrastructure	Provided by a third party or managed internally on a dedicated infrastructure	Data is stored and Managed on physical servers within organizations
Accessibility	Easily accessible over internet. Could have latency and control issues.	Easily accessible over internet. Latency could be improved.	Accessible primarily within the organization.
Cost	Less Costly. Depends on the vendor to vendor. Low initial investment	Comes with initial set-up costs. Can have better ROI over long term.	High initial costs to setup infrastructure. High cost of maintenance.
Scalability	Easy and quick	Not quick as compared to Public cloud	Limited scalability due to hardware
Deployment and maintenance	Can be deployed and easy to maintain.	Could be complex in managing and maintenance.	Requires in-house expertise.
Security	Could lead to security and compliance issues due to data saved in locations outside the organization.	Enhanced security and compliance and have greater control.	Complete control over data. Ideal for sensitive and private data.

Author: Ravi Akasapu

Comparison of Storage Options across Decision Factors



Public Cloud scores highest in Cost and Scalability but lower in Control and Security.

Private Cloud offers a balanced approach with moderate scores across all factors.

On-Premise scores highest in Control and Security but lower in Cost and Scalability.

categories	['Cost', 'Scalability', 'Security', 'Control', 'Maintenance', 'Compliance']
Public cloud scores	[5, 5, 3, 2, 4, 3]
Private cloud scores	[4, 4, 4, 4, 3, 4]
On-Premise scores	[2, 2, 5, 5, 3, 5]

Author: Ravi Akasapu

Integration

Source Systems, Meta-data Management, Reporting Tools, Governance, other Services

Author: Ravi Akasapu

Integration

Integration of various systems within a Data Architecture is critical in ensuring data consistency, reliability and accessibility. There could be several integrations that are required between various applications, source systems, and across platforms.

There are several key considerations that are need proper understanding

- **System Compatibility –** Software and Hardware compatibility along application compatibility
- **Data Interoperability –** All systems should be able to exchange data seamlessly
- **APIs –** Use of proper API's for data exchange. Depends on applications and types of data
- **Real-Time vs. Batch Processing –** Depends on use case as well as tools that handle the processing**.**
- **Security –** It is one of the important things while considering the data architecture. Need to look at Encryption, Access Management, Compliance, Regulatory requirements, Audit, Threat and Vulnerability and many more.
- **Scalability –** The system should be able allow Horizontal/vertical scaling as needed. Employing a modular approach can be useful in order to achieve proper scalability.

All the above considerations are applicable when connecting with Source systems, Storage solutions, Meta Data and Governance tools, API and other third party solutions.

Author: Ravi Akasapu

Impact Area	System Compatibility	Data Interoperability	APIs and Middleware	Real-time vs. Batch Processing	Security & Compliance	Scalability and Future Proofing
Source Systems	High	High	Medium	Medium	High	High
Metadata Management	Medium	High	Medium	Low	High	Medium
Reporting Tools	Medium	Medium	Medium	High	High	High
Governance	Medium	Medium	Low	Low	Very High	Medium
Other Services	High	High	Very High	High	High	Very High

Author: Ravi Akasapu

Data Categorization

Reporting, Active Staging, Legacy

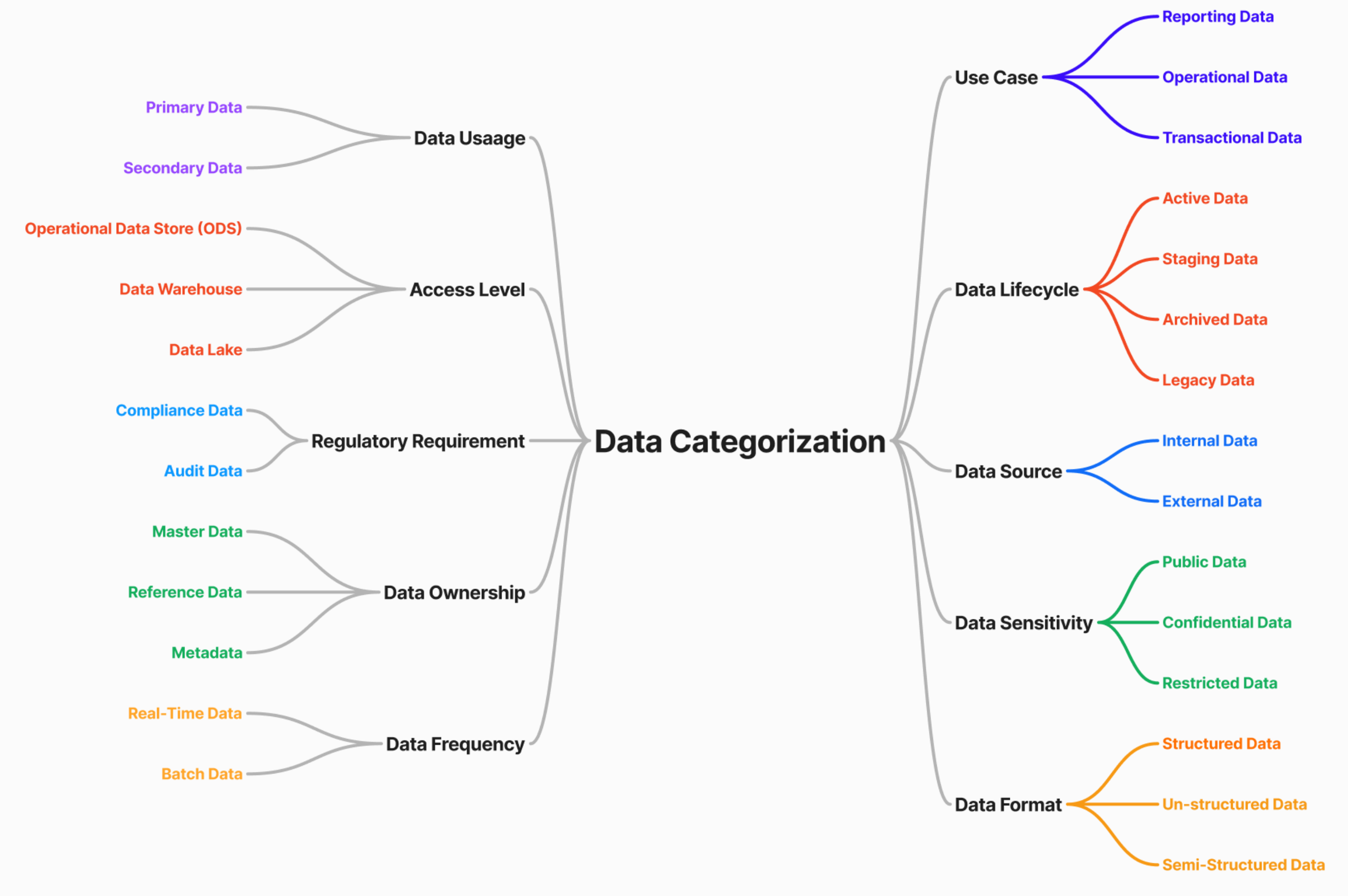
Author: Ravi Akasapu

Data Categorization

Data categorization is the process of classifying data into different categories based on its use, value, and lifecycle stage.

- Proper categorization of this data is essential for effective data management, storage, and retrieval.

- There can be many categories of the data depending on the kind of categorization.
  - **Data Usage**
  - **Access Level**
  - **Regulatory Requirements**
  - **Data Ownership**
  - **Data Frequency**
  - **Use Case**
  - **Data Lifecycle**
  - **Data Source**
  - **Data Sensitivity**
  - **Data Format**



Data Categorization

- Data Usaage
  - Primary Data
  - Secondary Data
- Access Level
  - Operational Data Store (ODS)
  - Data Warehouse
  - Data Lake
- Regulatory Requirement
  - Compliance Data
  - Audit Data
- Data Ownership
  - Master Data
  - Reference Data
  - Metadata
- Data Frequency
  - Real-Time Data
  - Batch Data
- Use Case
  - Reporting Data
  - Operational Data
  - Transactional Data
- Data Lifecycle
  - Active Data
  - Staging Data
  - Archived Data
  - Legacy Data
- Data Source
  - Internal Data
  - External Data
- Data Sensitivity
  - Public Data
  - Confidential Data
  - Restricted Data
- Data Format
  - Structured Data
  - Un-structured Data
  - Semi-Structured Data

Author: Ravi Akasapu

Other Considerations

Cost Considerations, Business and Data Teams,
Legacy systems

Author: Ravi Akasapu

Other Considerations

There could be many other factors that need to be considered in order to design a data platform. Below are some of the important factors to understand:

Cost Considerations

- Infrastructure
- Operational
- Scalability

Business & Data Teams

- Cross-Department Collaboration
- Skill Sets and Training
- Data Governance Framework

Legacy Systems

- Integration Challenges
- Cost-Benefit Analysis
- Data Quality Concerns

Author: Ravi Akasapu

Thank You !!!

Author: Ravi Akasapu