ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Framework of Hub and Spoke Data Governance Model for Cloud Computing

Ravi Kiran Koppichetti

USA

ABSTRACT

As data increasingly becomes a strategic asset, organizations encounter challenges managing and governing data effectively, especially in cloud environments. Large organizations with multiple business units worldwide often have a predominantly decentralized structure and a variety of data sources. This paper provides a framework for designing a practical data governance model for globally distributed organizations that use the hub-and-spoke model to run their business and operations strategically. Manufacturing, retail, logistics, and healthcare organizations are some of the few industries that gain great value with this framework. The framework addresses the growing need to manage distributed data silos while balancing centralized control with local flexibility.

The paper identifies five critical decision domains central to governance success and introduces the concepts of dataset rating and scope, which are vital for designing a robust data catalog. As organizations are rapidly expanding worldwide and adapting different cloud technologies and creating data silos, this framework provides a scalable solution for organizations to achieve governance objectives.

***Corresponding author**

Ravi Kiran Koppichetti, USA.

Received: March 08, 2023; **Accepted:** March 15, 2023; **Published:** March 22, 2023

Keywords: Data Governance, Cloud Computing, Framework, Data Management, Cloud Data Governance, Hub-and-Spoke Model

Introduction

Since the dawn of this century, companies have started finding value in data and treating it as a key strategic asset to gain sustainable strategic advantage. In the last two decades, successful companies have vastly invested in Cloud computing to increase productivity and implement the latest information systems. These investments have enabled companies to store, manage, and analyze data at unparalleled speed and low cost.

Like any other key strategic asset, data needs good governance to manage. Organizations in manufacturing, Retail, and other industries generate a lot of valuable and sensitive data that needs to be efficiently captured, stored, secured, and, if required, analyzed. A lack of trusted IT infrastructure to handle data adds risks and costs, impedes change with market needs, and leads to poor or wrong decisions.

Many organizations in Manufacturing and Retail have created and stored data in silos, and they seek to collect the data in a single system to manage and analyze data quickly. This data collection process increases the need for good governance principles and measures to protect sensitive data and minimize risk. The Hub and Spoke, data governance model, has emerged as a viable framework in response to cloud environments' data governance challenges and requirements.

In the Hub and Spoke model of cloud computing, the hub is the central authority that enforces governance policies, guarantees regulatory compliance, and manages shared tools and platforms across the spokes. The spoke in the organization can be defined as a business unit, set of projects, or even geographies. The spokes locally implement the policies enforced by the hub [1,2].

In this paper, I first discuss the main domains of Data Governance and how different levels of centralized, decentralized, and shared decision rights may be appropriate for various domains (spokes) in the Hub and Spoke Model of Data Governance.

Data Governance Models

The Data Governance model for cloud computing defines the structure, policy, and process an organization needs to use to manage the data assets effectively. The model is responsible for enforcing policies and ensuring compliance across all cloud components utilized by the organization. Two main traditional models (Centralized and Decentralized) were primarily used across all organizations before they became too complex and led to the origin of alternatives, such as Hub and Spoke and Data Mesh. In this paper, we will discuss the advantages and challenges of centralized, decentralized model and the reasons for origin of Hub and Spoke model.

Centralized Model

The central data team controls this model's data platforms, standards, policies, decisions, and compliance. This model ensures uniformity in policies and processes across the organization with apparent authority and accountability.

Although this model seems perfect, and many organizations have implemented it, a few significant problems exist. In this model, since all the data and platforms are maintained by the central data

team, bringing in data from new silos is a challenge. Also, since the data team doesn't necessarily have domain knowledge, it is difficult to derive value quickly from the data. In this model, all the data is centrally maintained in a single data warehouse, making exploring and analyzing data difficult [2]. Figure 1. Visualizes the Centralized Data Governance Model.

Figure 1: Centralized Model. Adapted from [3]

Decentralized Model

In this model, the business teams extract and analyze the data and build reports based on their business requirements instead of depending on the central data team. This approach helps business domains to respond to market trends and utilize the data quickly.

Like the centralized model, this model even has its issues. Since individual domains and users take ownership of data and data products, there is a lack of control over business definitions and struggles to maintain a single source of truth. It led to low confidence in data from senior leadership of the organizations. Secondly, since individual domains maintain ownership of data and data platforms, various competing products are used across the organization. It makes integrating data from multiple domains very difficult [2]. Figure 2. Visualizes the Decentralized Data Governance Model.

Figure 2: Decentralized Model. Adapted from [3]

Hub – Spoke Model

Extremely centralized or highly decentralized data models are not ideal for most organizations, and that led to the development of the Hub Spoke Model. This model balances centralized and decentralized models where the central data team (Hub) can control data quality and storage. At the same time, domain users (Spokes) can manipulate the data without breaching security. In this model, the central team (Hub) manages shared data assets and operations across the organization. Specific datasets or analyses can be shared using the Hub assets with the Spokes (company domains). Each spoke (domain) team specializes in a particular company domain and works with their stakeholders.

The Hub and Spoke model also allows scalability and flexibility for both Hub and Spoke data teams. Teams can quickly add or change new functions without impacting other parts of the business. Spoke teams have the flexibility to build their models and analyses. This flexibility allows the spoke teams to immediately modify models and data products to respond to changing business requirements.

Despite having so many benefits, there are also a few drawbacks that, if not acknowledged and carefully planned, can impact the business. In this model, the data is shared across spokes by the hub. So, the responsibility of data security and access control entirely lies on the Hub team and needs to be carefully managed. Additionally, if the Hub team manipulates the data or the logic, spoke teams should make necessary changes downstream to maintain data consistency. Figure 3. Visualizes the Hub and Spoke Data Governance Model.

Figure 3: Hub & Spoke Model. Adapted from [3]

Framework to Design Hub and Spoke Data Governance Model

Roles and Objectives

According to various knowledge bodies, the Data Governance framework specifies decision rights and accountability to ensure the appropriate behavior in creating, consuming, and controlling data and its analytics. The design framework presented in this paper aims to support consumers who need to design Hub and Spoke data governance using cloud computing services.

Industry associations and knowledge bodies like DAMA, IBM, and DGI have designed data governance frameworks. The framework provided by the Data Governance Institute consists of three components: rules and rules of engagement, people and organizational bodies, and processes [4]. Similarly, IBM's roadmap for effective data governance consists of fourteen steps [5]. Most of the components or steps mentioned in earlier frameworks cater highly to traditional IT systems.

Since this paper focuses on designing a hub-and-spoke framework for companies using cloud technologies, we will focus on the most relevant and practical components of the Hub-and-Spoke Model.

Establish Governance Objectives: In the Hub-and-Spoke Governance model, the primary objectives should be to centralize control and decision-making at the "Hub" while providing flexibility and local adaptation at the "Spokes" and defining and maintaining consistent standards across the company. Additionally, Governance should streamline communications among Hub and Spokes, leading to efficient operations, data sharing, reporting, and aligning local and company strategic interests.

Define Hub's Role: The "Hub" in the Hub and Spoke model of Data Governance plays a crucial role as the central point of control, creating and enforcing standards and policies across all spokes, optimizing resource allocation to minimize duplication of efforts, and aligning local initiatives with the company's interests. The Hub should create company-wide data quality, privacy, and security standards. It should deploy governance tools across the enterprise for metadata management, data cataloging, and data lineage. Hub should regularly monitor the spokes to ensure they adhere to governance policies through audits and dashboards.

Define Spoke's Role: The "Spoke" in the Hub and Spoke Governance model represents decentralized business units, projects, or departments responsible for implementing governance policies locally. Spoke should also be able to define and enforce additional policies to meet local data needs. It manages the data and data products specific to their domain and reports the governance metrics back to the Hub.

Governance Domains

Data governance refers to who has decision rights and is accountable for an organization's decision-making about its data assets [6]. This paper's Hub and Spoke Data Governance model framework includes these interrelated decision domains

- Data Principles
- Governance roles and responsibilities
- Data Catalog
- Data Quality
- Data Lifecycle

Data Principles: Practical data principles are crucial for aligning data management with business objectives. Organizations need to appoint a business owner for data assets to standardize the process of accountability and strategy alignment. Business owners and the central hub team bridge the centralized governance, and the individual business needs of each spoke.

The data principles will guide policy and standards development by defining data as an asset. Defining data as an asset enables organizations to share it among hubs and spokes to enhance operational efficiency and better decision-making. Since data is now being used for decision-making, business owners and data stewards will focus more on data consistency, integrity, and quality.

Compliance with regulatory requirements is vital to mitigate risks and to maintain the license to operate in many critical industries. The data principles guide how regulatory guidelines restrict or allow data usage to achieve specific strategic goals of the organization [6].

Governance Roles and Responsibilities: A Data Governance framework assigns clearly defined roles and responsibilities to ensure accountability and effective data management. This hub-and-spoke framework recognizes the requirement of Data owners, Analysts, Architects, and subject matter experts on spokes in the central hub team and the need for local data owners, analysts, and scientists in individual spokes. The data owners in the central hub and spokes should assign data stewards responsible for data quality, accuracy, and usage.

This paper on the Data Governance framework for cloud computing also recognizes the need for a cloud manager, provider, and broker. These are responsible for handling the cloud infrastructure and keeping it available for users across the organization.

Data Catalog: The Data Catalog acts as a centralized metadata repository for all the data owned and maintained by different spokes (Business domains). The metadata maintained in the data catalog contains data ownership details and descriptions that help in data discovery, access, and management while ensuring adherence to consistent governance policies set by the central and spoke teams.

The data catalog is an essential component in the framework as it enables users across the organization to find details of data owned by the hub and other spokes. It helps avoid data duplication and improves the usage of data for analytics. The metadata maintained in the data catalog should contain critical elements such as Data source and lineage, data quality metrics, Data owners, stewards, Access controls and usage policies, Business-specific definitions and annotations, and usage context in the Spoke.

Additionally, for the Hub and Spoke Data Governance model, Data Set Rating and Data Set Scope are two key dimensions of the Data Catalog.

Data Set Rating indicates the stringency followed in cleaning, normalizing, and generating the dataset. This framework categorizes Data Set ratings into three distinct types

- **Rating Type 1:** The data is raw or closest to raw form. This dataset required the lowest level of processing to make it available in a consumable format. It is not subjected to many cleansing and governance standards and can be used to answer some operational questions in the spokes.
- **Rating Type 2:** The data is cleansed but still has some errors and gaps. This dataset's format is consumable but may not have a clean and standard structure. Since the dataset is created after some level of processing, some documentation will be available. This dataset can be used to build reports to analyze or build a business case where limited errors are acceptable.
- **Rating Type 3:** The data is of high quality, monitored manually or using scripts regularly, and is available for consumption in a standard model suitable for reporting. This dataset can be used to generate performance, financial, and compliance reports.

The Data Set Scope indicates the intended audience of the dataset. This governance framework categorizes the scope into three types

- **Scope Type 1:** The data is local and sourced from a specific organization's Spoke (domain). The dataset can only be used or shared within the same domain. This dataset is local to a single spoke.
- **Scope Type 2:** The data is sourced from more than one spoke (domain) or can be shared across domains. This dataset is used across spokes.
- **Scope Type 3:** The data is sourced from different spokes (domains) or can be shared across multiple domains. This dataset is sourced or used across the organization.

Data Quality: Like product quality, poor data quality affects the ability to use data [6]. It can impact organizations' decision-making at both the operational and strategic levels. In the Hub and Spoke model of Data Governance, the central hub team is primarily responsible for Data Quality. They are responsible for standardizing, validating, and cleansing the data before it is available for usage across spokes (domains).

Although the central team has primary responsibility, if a dataset is produced, owned, and used within a spoke (domain), then the data owner and stewards are responsible for maintaining data quality as per the standards set by the governance principle.

The data quality has many components, and this framework recognizes the following to be relevant

- **Data Accuracy:** The central data hub team primarily maintains data accuracy and ensures that it is standardized, cleansed, and validated before making it available for all users across the organization.
- **Timeliness:** The central data hub team ensures that the data available is updated regularly. This component allows the spokes (domains) to access the most recent information for analysis and decision-making when necessary.

Data Lifecycle: The Data Lifecycle is a sequence of stages a dataset goes through until it perishes. It discusses how a dataset is created, collected, maintained, analyzed, archived, and finally deleted [6].

In the hub-and-spoke model, data is created at the spokes. The central data team at the hub then ingests, transforms, collects, and models this data. After the hub team makes the dataset available for consumption by the spokes, the spokes analyze the data for their specific needs [7].

Conclusion

This paper describes a framework for designing a Hub-and-Spoke Data Governance Model for globally operating organizations that maintain data, applications, data products, and necessary tools on the cloud. First, I introduced the concept of data governance and prominent data governance models. Second, I established the governance objectives and the role of "Hub" and "Spoke." Lastly, I introduced the five most relevant decision domains and discussed them in detail. I have also introduced the dataset rating and dataset scope, which play a crucial role in designing the data catalog for the Hub and Spoke model. The paper presents an initial attempt to illustrate a conceptual framework for designing a Hub and Spoke Data Governance model for organizations using cloud computing. As more and more companies continue moving to cloud computing, they enable great power to Spoke (Domains) using modern cloud-based integration tools. This transition will force organizations to adapt to the hub and spoke model. The framework discussed in this paper may provide helpful guidance for organizations.

References

1. Niemi E (2011) Designing a data governance framework. Proceedings of the IRIS Conference, Oslo, Norway. Available at: https://research.aalto.fi/en/publications/designing-a-data-governance-framework.
2. King T (2022) The Data Governance Hub and Spoke Model: Why It Works. Solutions Review. Available at: https://solutionsreview.com/data-management/the-data-governance-hub-and-spoke-model-why-it-works/.
3. Ironside Group (2023) Learn why so many companies are turning to a hub-and-spoke data model. Ironside Group Blog. Available at: https://www.ironsidegroup.com/blog/learn-why-so-many-companies-are-turning-toa-hub-and-spoke-data-model/.
4. The Data Governance Institute (2015) Definitions of Data Governance. Available at: http://www.datagovernance.com/adg_data_governance_definition/.
5. IBM Institute for Business Value and IBM Strategy and Change (2007) The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance. Gov An Int J Policy Adm pp: 1-16.
6. Khatri V, Brown CV (2010) Designing data governance. Communications of the ACM 1: 148-152.
7. Al-Ruithe M, Benkhelifa E, Hameed K (2016) A conceptual framework for designing data governance for cloud computing. Procedia Computer Science 94: 160-167.