International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Data Mesh Architecture Support in Snowflake vs Databricks: A Comprehensive Comparison

Rameshbabu Lakshmanasamy

Senior Data Engineer, Jewelers Mutual Group

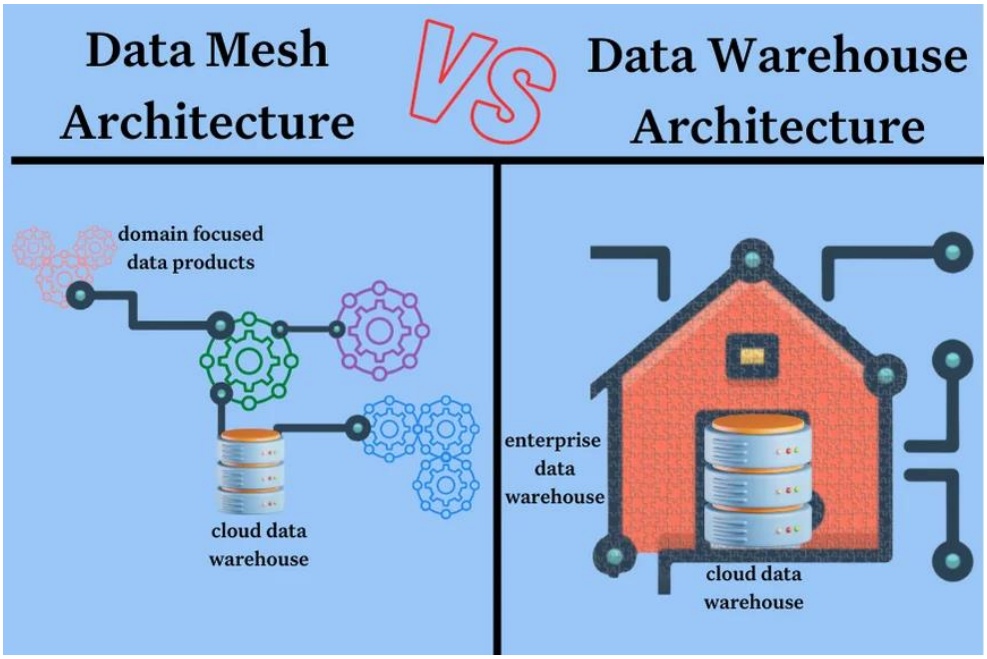
Abstract: ***In current world, where data is considered new oil, organizations face numerous challenges, due to increased complexity and volume of data as they require them for centralized data solutions, decision - making, including data lakes and data warehouses, do not effectively fulfill the needs of real - time data and faster data processing. As a response to these challenges, Data Mesh Architecture has proferred a shift from traditional data management system by adopting a decentralized data management and usage system in large - scale enterprises.***

Keywords: Data Mesh Architecture, Data products, Decentralized data management, Snowflake, Databricks

1. Introduction

Data Mesh is built on four core principles: domain - specific data decentralization, data as a product, self - serve data infrastructure, and federated computational control. This model prescribes that every domain or business line within a firm owns its data to support greater flexibility, modularity, and governance. Data Mesh decentralizes data management and democratizes access, thus enabling orgs to escape from bottlenecks occasioned by centralization teams (Armbrust et al., 2021). This leads to better data quality because of less processing time.

This paper aims to identify similarities and differences in support of implementing Data Mesh architecture by two significant cloud - based platforms: Snowflake and Databricks. Snowflake is a cloud data warehousing company that has impressive scalability. In contrast, Databricks has a unified analytics platform based on Spark and Delta Lake that corresponds to main principles of Data Mesh. This paper shall discuss what the platforms facilitate in building data products, dealing with data contracts, and creating scalability, governance, and performance solutions in decentralized data contexts.



Understanding Data Mesh Architecture:
Since organizations grow and acquire a large amount of data, traditional centralized data governance models cannot satisfy the increasing requirements for quick adaptation, instantaneous analysis, and effective data utilization. Data Mesh has become an innovative solution that opens up data ownership and management, shortcomings observed in centralized architecture such as data lakes and warehouses. It is based on four foundational principles: The four concepts that define this space are domain - oriented decentralized data ownership, data as a product, self - serve data infrastructure, and federated computational governance (Armbrust et al., 2021). These principles transfer the key from large infrastructures to a more flexible, domain - oriented architecture, which creates more possibilities for effectively storing and using data.

The Decentralized Data Ownership within Domain Orientation means that the data is owned by the business units who are to be responsible for creation & processing of such

data. This decentralizes data analysis for the teams using the data, removing inefficiencies and bottlenecks caused by a large centralized data team. Data as a product, domain teams engage with their data in the same way that is expected of tangible products and their deliverers, with due consideration of the quality of the final product and how easy it is to find and use it (Goedegebuure et al., 2023). To facilitate this, self - service data platforms allow teams to create, govern and deploy their data assets autonomously, minimizing technology requests on central IT function and innovation. However, federated computationally decentralized governance also enforces global governance policies of security, compliance, and data quality across all teams, achieving balanced computation and organization.

For instance, data lakes and data warehouses have limitations in scaling, flexibly accommodating more data, and management as organizations evolve. They have centralized data management, which results in some data latency, meaning data that requires the approval of a distinct team before it can be released. These architectures also need help with scalability especially if dealing with real - time or if there is some domain case processing by these systems; this often leads to added cost and system complexity. These limitations are addressed through adopting Data Mesh due to characterized decentralization and enhanced data ownership, allowing for scalability and agility.

Snowflake: Support for Data Mesh Architecture:
As stated in the previous section, Snowflake is an online cloud - based data platform that aims to support properly standardized data warehousing solutions while shaping itself to be scalable, safe, and high - performance, categorized on critical components which include compute - isolated from the storage, automatic scaling and data sharing with security in its core agenda. Such separation results in organisations' ability to scale resources separately or in harmony to enhance performance and minimize cost (Snowflakes, 2019). The flexibility of Snowflake makes it possible to address large amounts of workload due to its scalability; on the other hand, the firm's security and capability to share data is beneficial for business enterprises that need to securely share data in different domains.

According to the concept of Data Mesh, controlling procedures are decentralized because each domain gains independence in managing its virtual warehouse supported by Snowflake. It also enables sector - specific data goods, for the same reason that teams can develop, contain, and safely share data commodities using the data - sharing provision of Snowflake. Also, its self - service features support the computing resources set by different working teams to enhance the creativity of innovative procedures. Federated computational governance is achieved with centralized governance control over different structures through the Federated RBAC Model while preserving domain - level autonomy and secure compliance.

Snowflake's strengths are its scalability, Data sharing, and simplicity of infrastructure, so domain teams do not necessarily need to be experts in computer science. Nevertheless, Snowflake is more inclined toward batch processing instead of real - time data processing, and some organizations may face problems in elaborating workloads for certain domain - driven needs in more intricate settings (Snowflakes, 2019).

Databricks: Support for Data Mesh Architecture:
Databricks is an integrated data management tool based on Apache Spark, focused mainly on big data processing and machine learning (Databrick, 2022). The Delta Lake enriches the traditional data lake structures through Transaction capability, Schema validation, and time integration. Regarding a Technical Data Stack, Databricks deals exceptionally well with integrating with the major cloud providers for real - time data processing and analytics with proper infrastructure. This makes Databricks flexible for organizations that practice Data Mesh architecture, where domain - specific data processing pipelines and sovereign data ownership are valued.

Following the Data Mesh approach, Databricks enables teams to build their data assets with Delta Lake, providing transactions and scalability with ACID support. Interactions across various domain teams can be easily managed with tools such as the Databricks Workspace and MLflow where Data engineers and Data Scientists can work in harmony (Sharma, 2024). Thus, Databricks enables the self - serve data infrastructure approach by providing interactive notebooks, data pipelines, and simple UIs for building and maintaining data products. There are also rich governance features: schema enforcement, role - based security access, and connections to third - party compliance tools.

Databricks is great at real - time processing and machine learning applications, so it will perform well in domains that use predictive modelling or real - time decision - making. Hadoop reflects its deep integration with Apache Spark and support for streaming data, which allows it to have a competitive advantage over other platforms; however, the platform's management might be challenging compared to more SQL - oriented solutions (Databrick, 2022). Databricks is very flexible and designed to deal with structured and unstructured data, but some companies that only require SQL querying and structured data handling the Databricks platform might be a bit complex.

Snowflake and Databricks: Better Suited for Decentralization:
Snowflake and Databricks are well - suited to support the decentralized nature of Data Mesh compared to traditional data architectures. Snowflake's architecture, separated compute and storage, elastic scaling, and secure data sharing, allows domain teams to create and manage their data products independently (Vinnikainen, 2023). Snowflake's strong governance and compliance features ensure decentralized teams can adhere to global data governance standards.

Databricks, built on Apache Spark and Delta Lake, offers a more flexible and scalable platform for managing both batch and real - time data products. With its support for ACID transactions, schema enforcement, and real - time streaming, Databricks enables domain teams to build robust, reliable data products in a decentralized environment while maintaining consistency and data quality (Eguíluz Romero, 2023).

Capabilities	databricks	snowflake
Services	Cloud Data Warehouse, Data Science Pipelines/Machine Learning	Cloud Data Warehouse, Data Science Pipelines/Machine Learning
Scalability	Auto-scaling	Auto-scaling with 128 node limitation
Data Structures	All data types (raw, audio, video, files, structured, semi-structured, etc.)	Semi-structured or Structured
User Friendliness	Some learning curve, with more coding but more depth & capability	Easier UI for data warehouse, but less depth & capability across the board
Querying	SQL, Spark Dataframe, Koalas	SQL
Performance	Batch or Streaming	Batch-based
Data Science & Machine Learning	Built-in ecosystem	Not native

Scalability and Performance:
Snowflake and Databricks have similar levels of scalability and performance; however, they are designed for different types of workloads and use cases, meaning each platform is better suited for other tasks.

Snowflake: The architecture of Snowflake is released from any hardware constraints and is optimized for extensibility of both compute and storage resources. From there, Snowflake is especially suited for querying big structured datasets and processing intricate analytical queries. Snowflake thrives on batch processing and providing a high tempo of query functionality for multiple teams without slowing down. Its ability to handle concurrent workloads from various clusters through a multi - cluster architecture also makes it suitable for organizations that perform a lot of data querying (Eguíluz Romero, 2023). Designing Snowflake works in a way that, as data scales, there is no scalability issue, as demonstrated by the fact that users only pay for storage consumed plus compute service.

Databricks: In contrast, Databricks is highly suitable for streaming or real - time analytics as well as machine learning jobs. Set up on the Apache Spark protocol, Databricks has better throughput for integrating and manipulating structured and unstructured data, more significantly recognized in the processing and containing of large - sized data change and intricate analytics streams. The stream data processing capability of Databricks makes it a good fit for real - time processing scenarios, including detection of safety violations or device maintenance (L'Esteve, 2023). Delta Lake, as its data architecture, makes data reliable and performant through, tournament ACID which ensures consistency in big data environments. It is also very efficient, especially for machine learning and AI purposes, because of its perfect integration with data science and Machine learning tools.

Governance and Security:
Governance and security are fundamental in any data architecture, particularly when applied to the decentralized nature of Data Mesh. Both Snowflake and Databricks provide the users with detailed governance guides in these areas, though they might approach these questions differently.

Snowflake: Snowflake's governance model is structured around centralized security characteristics: role - based access controls (RBAC), data encryption, data masking, and auditing. These features enable organizations to achieve decentralized security and compliance of the decentralized teams and the global governance benchmarks (Eguíluz Romero, 2023). The security controls are one of Snowflake's greatest assets, especially for industries or products that must address GDPR/CCPA/HIPAA regulations. In the context of Data Mesh, each domain can organize its data as it wishes but with a specific, optimal set of rules for the organization. This is what Snowflake offers with federated governance.

Databricks: Another critical area well addressed in Databricks is governance with the fine - grained access controls, data lineage tracking, and version control achieved through Delta Lake. Critically important is the real - time governance capability, and therefore, the platform is particularly effective for decentralized systems, where the data is often changing or provided in real - time mode (Armbrust et al., 2021). Databricks works well with other governance outside tools and the framework, enabling the organization to have federated governance for the data.

2. Conclusion

In the case of comparing Snowflake and Databricks for utilizing Data Mesh architecture, it is meaningful to state that both platforms have strengths that can satisfy the organization's needs. Snowflake is uniquely suitable for managing highly standardized data at an industrial level, owing to the following aspects: fully elastic, separating computing and storage, and governance. That is why it is most beneficial to organizations using SQL - based querying and must maintain strict governance in a decentralized structure. On the other hand, Databricks naturally performs well in scenarios that need real - time analytics, machine learning workflows, and if it is unstructured or semi - structured data processing. For domains related to data science and advanced analytics, architecture Delta Lake with support for ACID transaction logs and real - time streaming adds a lot of versatility.

References

[1] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (Vol.8, p.28). https:

//15721. courses. cs. cmu. edu/spring2023/papers/02 - modern/armbrust - cidr21. pdf
[2] Databricks. (2022). Building a Data Mesh Based on the Databricks Lakehouse, Part 2. Databricks. https: //www.databricks. com/blog/building - data - mesh - based - databricks - lakehouse - part - 2
[3] Eguíluz Romero, J. (2023). Sustainable Data Warehousing for Businesses: Comparison of Data Vault 2.0 vs. Traditional Data Modeling Methods in the Data Engineering Sphere. https: //www.diva - portal. org/smash/record. jsf?pid=diva2: 1826876
[4] Goedegebuure, A., Kumara, I., Driessen, S., VAN DEN HEUVEL, W. J., Monsieur, G., Tamburri, D. A., & Nucci, D. (2023). Data mesh: a systematic gray literature review. ACM Computing Surveys. https: //dl. acm. org/doi/abs/10.1145/3687301
[5] L'Esteve, R. C. (2023). Decentralizing Data and Democratizing Analytics. In The Cloud Leader's Handbook: Strategically Innovate, Transform, and Scale Organizations (pp.79 - 104). Berkeley, CA: Apress. https: //link. springer. com/chapter/10.1007/978 - 1 - 4842 - 9526 - 7_6
[6] Sharma, R. (2024, August 20). Databricks vs. Snowflake: A Comprehensive Comparison for Data - Driven Decision - Making. Sprinkledata. com; Sprinkle Data. https: //www.sprinkledata. com/blogs/databricks - vs - snowflake - a - comprehensive - comparison - for - data - driven - decision - making
[7] Snowflakes. (2019). Data Mesh for Self - Service Data. Snowflake. https: //www.snowflake. com/guides/data - mesh - self - service - data/
[8] Solanki, J. (2023, March 17). Snowflake vs Databricks | Comparison & Features. Decube. io; Decube Inc. https: //www.decube. io/post/snowflake - vs - databricks
[9] Vinnikainen, O. (2023). Data mesh: a holistic examination of its principles, practices, and potential. https: //lutpub. lut. fi/handle/10024/166353