การเปรียบเทียบ Google Dataplex และ Microsoft Purview สำหรับวิศวกรข้อมูล

บทสรุปสำหรับผู้บริหาร

ในยุคที่ข้อมูลมีการกระจายตัวและมีความซับซ้อนสูง การกำกับดูแลข้อมูล (Data Governance) ได้กลายเป็นรากฐานสำคัญสำหรับการขับเคลื่อนคุณค่าทางธุรกิจและการนำ AI มาใช้ Google Dataplex และ Microsoft Purview เป็นสองแพลตฟอร์มชั้นนำที่ออกแบบมาเพื่อจัดการกับความท้าทายเหล่านี้ โดยมีจุดแข็งและแนวทางที่แตกต่างกัน

Google Dataplex มุ่งเน้นไปที่การสร้าง "Data Fabric" ที่เป็นหนึ่งเดียวสำหรับการกำกับดูแลข้อมูลและสินทรัพย์ AI ภายในระบบนิเวศของ Google Cloud โดยเฉพาะ แพลตฟอร์มนี้โดดเด่นในการผสานรวมอย่างลึกซึ้งกับบริการของ Google Cloud การสนับสนุนสถาปัตยกรรม Data Mesh และความสามารถด้าน AI/ML สำหรับการจัดการเมทาดาตาและข้อมูลเชิงลึก

ในทางตรงกันข้าม Microsoft Purview นำเสนอชุดโซลูชันที่ครอบคลุมทั้งการกำกับดูแลข้อมูล ความปลอดภัยของข้อมูล และการจัดการความเสี่ยงและการปฏิบัติตามข้อกำหนด โดยเน้นย้ำถึงแนวคิด "Data as a Product" และการกำกับดูแลแบบรวมศูนย์ (Federated Governance) การผสานรวมอย่างแน่นแฟ้นกับบริการของ Azure และ Microsoft 365 เป็นจุดแข็งหลักของ Purview

การตัดสินใจเลือกระหว่าง Dataplex และ Purview ขึ้นอยู่กับกลยุทธ์คลาวด์หลักขององค์กร สถาปัตยกรรมข้อมูลที่ต้องการ และลำดับความสำคัญเฉพาะด้าน เช่น ความต้องการด้านความปลอดภัยและการปฏิบัติตามข้อกำหนดที่เข้มงวด หรือการมุ่งเน้นไปที่การสร้าง Data Mesh ภายในคลาวด์เดียว

บทนำสู่การกำกับดูแลข้อมูลบนคลาวด์

ภูมิทัศน์ข้อมูลในปัจจุบันมีการเปลี่ยนแปลงอย่างรวดเร็ว โดยข้อมูลมีการกระจายตัวอย่างมหาศาลข้ามสภาพแวดล้อมแบบไฮบริด มัลติคลาวด์ และแอปพลิเคชัน SaaS [1] การกระจายตัวของข้อมูลนี้สร้าง

ความซับซ้อนอย่างมากในการบริหารจัดการ การรับรองคุณภาพ และการรักษาการปฏิบัติตามข้อกำหนด ซึ่งเป็นกระบวนการที่ใช้เวลาและซับซ้อน [6] เพื่อตอบสนองต่อความท้าทายนี้ แพลตฟอร์มการกำกับดูแลข้อมูลแบบครบวงจรจึงมีความจำเป็นอย่างยิ่งในการรวมการกำกับดูแล ความปลอดภัย และการปฏิบัติตามข้อกำหนดเข้าไว้ในระบบอัจฉริยะเดียว [3]

การกำกับดูแลข้อมูลได้พัฒนาจากการเป็นเพียงกลไกป้องกันที่ขับเคลื่อนด้วยการปฏิบัติตามข้อกำหนดไปสู่บทบาทเชิงกลยุทธ์ในการปลดล็อกคุณค่าทางธุรกิจ ขับเคลื่อนนวัตกรรมอย่างมีความรับผิดชอบ และสนับสนุนการวิเคราะห์ขั้นสูง รวมถึงเวิร์กโหลด AI/ML [2] การเปลี่ยนแปลงนี้แสดงให้เห็นว่าการกำกับดูแลข้อมูลไม่ได้เป็นเพียงแค่การทำเครื่องหมายในช่องสี่เหลี่ยมตามข้อบังคับอีกต่อไป [3] แต่เป็นสิ่งสำคัญอย่างยิ่งในการรับรองความถูกต้อง ความน่าเชื่อถือ และความสามารถในการใช้งานของข้อมูลเพื่อวัตถุประสงค์ทางธุรกิจที่สำคัญ

สำหรับวิศวกรข้อมูล ซึ่งเป็นผู้ที่อยู่แนวหน้าในการออกแบบ สร้าง และบำรุงรักษาไปป์ไลน์ข้อมูลและโครงสร้างพื้นฐาน การมีโซลูชันการกำกับดูแลข้อมูลที่แข็งแกร่งเป็นสิ่งจำเป็นอย่างยิ่ง แพลตฟอร์มเหล่านี้ช่วยให้มั่นใจว่าข้อมูลสามารถค้นพบได้ ถูกต้อง น่าเชื่อถือ และได้รับการปกป้อง [5] การกำกับดูแลที่มีประสิทธิภาพช่วยลดภาระการดำเนินงาน ปรับปรุงคุณภาพข้อมูล อำนวยความสะดวกในการแบ่งปันข้อมูล และรับรองการปฏิบัติตามข้อกำหนด ซึ่งส่งผลโดยตรงต่อประสิทธิภาพและความน่าเชื่อถือของเวิร์กโฟลว์วิศวกรรมข้อมูล [2] ความต้องการแพลตฟอร์มแบบครบวงจรจึงเกิดขึ้นจากการที่ข้อมูลในปัจจุบันมีการกระจายตัว ทำให้การบริหารจัดการข้อมูลเป็นเรื่องที่ซับซ้อนและใช้เวลานาน หากไม่มีแพลตฟอร์มเหล่านี้ วิศวกรข้อมูลจะต้องเผชิญกับความท้าทายในการควบคุมและมองเห็นข้อมูลที่กระจัดกระจายอยู่ในไซโลต่างๆ ซึ่งจะเพิ่มภาระการดำเนินงานและความเสี่ยงโดยไม่จำเป็น

Google Dataplex: มุมมองของวิศวกรข้อมูล

Google Dataplex ได้รับการออกแบบมาเพื่อทำหน้าที่เป็น "Data Fabric" อัจฉริยะที่รวมศูนย์การจัดการและการกำกับดูแลข้อมูลทั่วทั้งองค์กร [12]

ความสามารถหลักและคุณสมบัติ

แค็ตตาล็อกเมทาดาตาแบบรวมศูนย์และการค้นพบข้อมูล

Dataplex Universal Catalog ทำหน้าที่เป็นโซลูชันการกำกับดูแลที่ชาญฉลาดและเป็นหนึ่งเดียวสำหรับสินทรัพย์ข้อมูลและ AI ใน Google Cloud โดยมีรายการข้อมูลส่วนกลางที่รวบรวมเมทาดาตาทางธุรกิจ เทคนิค และรันไทม์ทั้งหมด [6] แพลตฟอร์มนี้ใช้ประโยชน์จากปัญญาประดิษฐ์ (AI) และการเรียนรู้ของเครื่อง (ML) เพื่อค้นหาความสัมพันธ์และความหมายในเมทาดาตา ซึ่งช่วยให้การสอบถามข้อมูลและข้อมูลเชิงลึกง่ายขึ้น [6]

สำหรับวิศวกรข้อมูล Dataplex ช่วยให้สามารถดึงเมทาดาตาสำหรับทรัพยากร Google Cloud ที่หลากหลาย เช่น BigQuery, Cloud SQL, Spanner, Vertex AI, Pub/Sub, Dataform และ Dataproc Metastore [6] นอกจากนี้ยังรองรับการสแกนข้อมูลทั้งแบบมีโครงสร้างและไม่มีโครงสร้างใน Cloud Storage buckets เพื่อแยกและจัดทำแค็ตตาล็อกเมทาดาตา [6] ความสามารถในการค้นหาข้อมูลของ Dataplex ยังรวมถึง "Data Insights" ที่ขับเคลื่อนด้วย AI ซึ่งสร้างคำถามภาษาธรรมชาติเกี่ยวกับข้อมูล และ "Semantic metadata search" ที่ช่วยให้ผู้ใช้ค้นหาข้อมูลโดยใช้ภาษาธรรมชาติ [17]

การเน้นการรวบรวมเมทาดาตาทางธุรกิจ เทคนิค และรันไทม์ และการใช้ AI เพื่อค้นหาความสัมพันธ์แสดงให้เห็นว่า Dataplex ถือว่าเมทาดาตาเป็นชั้นที่ใช้งานได้และชาญฉลาดสำหรับการกำกับดูแล การนำเสนอการค้นหาด้วยภาษาธรรมชาติผ่าน AI ช่วยให้การค้นพบข้อมูลเป็นประชาธิปไตยมากขึ้น สิ่งนี้หมายความว่าวิศวกรข้อมูลจะต้องพิจารณาการสร้างและคุณภาพของเมทาดาตาเป็นส่วนสำคัญของการออกแบบไปป์ไลน์ข้อมูลของตน หากเมทาดาตาพื้นฐานไม่มีคุณภาพ การค้นหาและข้อมูลเชิงลึกที่ขับเคลื่อนด้วย AI จะไม่มีประสิทธิภาพ ซึ่งนำไปสู่ความไม่ไว้วางใจในแค็ตตาล็อก อย่างไรก็ตาม การที่ AI เข้ามาช่วยในส่วนนี้อาจลดคำขอสอบถามเฉพาะกิจสำหรับการทำความเข้าใจข้อมูลพื้นฐานลงได้ ทำให้วิศวกรมีเวลาไปทำงานที่ซับซ้อนมากขึ้น

คุณภาพข้อมูลและโปรไฟล์ข้อมูลอัตโนมัติ

Dataplex ช่วยให้สามารถกำหนดและวัดคุณภาพของข้อมูลในตาราง BigQuery โดยตรวจสอบข้อมูลตามนโยบายขององค์กรและบันทึกการแจ้งเตือนหากข้อมูลไม่เป็นไปตามเกณฑ์คุณภาพที่กำหนด [6] การทำโปรไฟล์ข้อมูลจะระบุลักษณะทั่วไปของข้อมูลในคอลัมน์ (เช่น ค่าข้อมูลทั่วไป การกระจายข้อมูล และจำนวนค่าว่าง) ซึ่งเป็นสิ่งสำคัญสำหรับการจัดประเภทข้อมูลและการประกันคุณภาพ [6]

กฎคุณภาพข้อมูลสามารถจัดการได้ "ในรูปแบบโค้ด" โดยมีตัวเลือกในการแนะนำกฎตามผลลัพธ์การสแกนโปรไฟล์ข้อมูลของ Dataplex Universal Catalog รวมถึงกฎที่กำหนดไว้ล่วงหน้า (ระดับแถว, ระดับรวม) และกฎ SQL แบบกำหนดเอง [6] ประเภทกฎที่กำหนดไว้ล่วงหน้า ได้แก่

RangeExpectation, NonNullExpectation, SetExpectation, RegexExpectation, Uniqueness และ StatisticRangeExpectation [19] กฎ SQL แบบกำหนดเองรองรับเงื่อนไขระดับแถว เงื่อนไขระดับตาราง (SQL รวม) และการยืนยัน SQL [6] กฎเหล่านี้สามารถเชื่อมโยงกับมิติข้อมูล

เช่น ความสดใหม่ ปริมาณ ความสมบูรณ์ ความถูกต้อง ความสอดคล้อง ความแม่นยำ และความเป็นเอกลักษณ์ เพื่อผลลัพธ์ที่รวมกัน [6] ผลการสแกนยังสามารถส่งออกไปยัง BigQuery เพื่อการวิเคราะห์เพิ่มเติมและการจัดทำรายงานแบบกำหนดเอง [6]

ความสามารถในการจัดการกฎคุณภาพข้อมูลและการปรับใช้ในรูปแบบโค้ดนั้นสอดคล้องโดยตรงกับแนวทางปฏิบัติของ DataOps และ DevOps สมัยใหม่ ซึ่งช่วยให้วิศวกรข้อมูลสามารถควบคุมเวอร์ชัน ทดสอบ และปรับใช้การตรวจสอบคุณภาพข้อมูลโดยอัตโนมัติพร้อมกับไปป์ไลน์ข้อมูลของตน การแนะนำกฎตามการสแกนโปรไฟล์ข้อมูลยังช่วยลดความยุ่งยากในการตั้งค่ากฎคุณภาพข้อมูลในเบื้องต้น สิ่งนี้บ่งชี้ถึงการเปลี่ยนแปลงที่สำคัญจากสคริปต์การตรวจสอบข้อมูลด้วยตนเองแบบตอบโต้ ไปสู่แนวทางที่เน้นการประกาศ ใช้โค้ด และเป็นอัตโนมัติ วิศวกรข้อมูลสามารถฝังคุณภาพข้อมูลโดยตรงในไปป์ไลน์ CI/CD ของตน ซึ่งช่วยปรับปรุงความสมบูรณ์ของการผลิตข้อมูลและลดความพยายามด้วยตนเองในการระบุและแก้ไขปัญหาข้อมูล การควบคุมที่ละเอียดอ่อนที่นำเสนอโดยกฎที่กำหนดไว้ล่วงหน้าและกฎ SQL แบบกำหนดเองช่วยให้วิศวกรสามารถใช้ตรรกะการตรวจสอบที่ซับซ้อนและเฉพาะเจาะจงทางธุรกิจได้โดยตรงภายในแพลตฟอร์มการกำกับดูแล เพื่อให้มั่นใจในความน่าเชื่อถือของข้อมูลในวงกว้าง

อย่างไรก็ตาม การตรวจสอบคุณภาพข้อมูลและการทำโปรไฟล์ข้อมูลจัดอยู่ในหมวดหมู่ "Premium Processing" ของ Dataplex ซึ่งมีค่าใช้จ่ายที่วัดเป็น Data Compute Unit (DCU) ชั่วโมง และไม่มีระดับฟรี [12] เอกสารแนะนำให้ "เลือกใช้คุณสมบัติอย่างรอบคอบ" "ใช้การสุ่มตัวอย่างและขอบเขตแบบเพิ่มขึ้น" และ "ลดความซับซ้อนของกฎคุณภาพข้อมูล" เพื่อ "ควบคุมการใช้งาน" [12] นี่แสดงว่าการเปิดใช้งานการทำโปรไฟล์และตรวจสอบคุณภาพข้อมูลอย่างครอบคลุมโดยไม่มีการวางแผน อาจนำไปสู่ค่าใช้จ่ายที่สูงและไม่คาดคิด วิศวกรข้อมูลจึงต้องมีส่วนร่วมอย่างแข็งขันในกลยุทธ์การเพิ่มประสิทธิภาพต้นทุน เช่น การทำโปรไฟล์เฉพาะชุดข้อมูลที่สำคัญ การใช้การสแกนแบบเพิ่มขึ้นสำหรับตารางที่อัปเดตบ่อยครั้ง และการออกแบบกฎคุณภาพข้อมูลอย่างมีประสิทธิภาพ สิ่งนี้ทำให้วิศวกรต้องเข้าใจแบบจำลองราคาและผลกระทบโดยตรงต่อการเลือกสถาปัตยกรรมและการดำเนินงาน โดยต้องสร้างสมดุลระหว่างความน่าเชื่อถือของข้อมูลและข้อจำกัดด้านงบประมาณ

สายข้อมูลแบบครบวงจร (End-to-End Data Lineage)

Dataplex ช่วยให้สามารถติดตามการเคลื่อนที่ของข้อมูลผ่านระบบต่างๆ รวมถึงแหล่งที่มา ปลายทาง และการแปลงที่ใช้ [6] มีการจัดเตรียมสายข้อมูลอัตโนมัติสำหรับแหล่งข้อมูล Google Cloud ที่หลากหลาย เช่น BigQuery, Cloud Data Fusion, Cloud Composer, Dataflow, Dataproc และ Vertex AI [6]

แพลตฟอร์มนี้ยังรองรับการขยายไปยังแหล่งข้อมูลบุคคลที่สามและระบบภายนอกผ่าน Data Lineage API และการผสานรวม OpenLineage ซึ่งช่วยให้สามารถบันทึกข้อมูลสายข้อมูลด้วยตนเองได้ [6] ข้อมูลสายข้อมูลถูกจัดระเบียบโดยใช้โมเดลลำดับชั้นของกระบวนการ การรัน และเหตุการณ์ และ

สามารถแสดงเป็นกราฟ การแสดงภาพพาธ หรือในมุมมองรายการ [6]

ข้อจำกัดที่สำคัญคือข้อมูลสายข้อมูลทั้งหมดจะถูกเก็บไว้ในระบบเพียง 30 วันเท่านั้น [20] นี่เป็นข้อจำกัดที่ส่งผลกระทบโดยตรงต่อกลยุทธ์การกำกับดูแลข้อมูลระยะยาว สำหรับองค์กรที่มีข้อกำหนดการปฏิบัติตามกฎระเบียบที่เข้มงวด (เช่น GDPR, HIPAA, SOX) หรือความต้องการในการตรวจสอบภายในที่ต้องการแหล่งที่มาของข้อมูลนานกว่า 30 วัน สายข้อมูลใน Dataplex ไม่เพียงพอในฐานะแหล่งความจริงเพียงแหล่งเดียว วิศวกรข้อมูลจะต้องออกแบบและใช้งานกลไกภายนอกเพิ่มเติมเพื่อดึงและจัดเก็บเมทาดาตาของสายข้อมูล (เช่น การใช้ Data Lineage API เพื่อส่งออกข้อมูลไปยังตาราง BigQuery หรือที่เก็บข้อมูลถาวรอื่นๆ) สิ่งนี้เพิ่มความซับซ้อนทางสถาปัตยกรรม ความพยายามในการพัฒนา และค่าใช้จ่ายในการดำเนินงานอย่างต่อเนื่อง ทำให้คุณสมบัติที่ดูเหมือนจะรวมอยู่ในตัวกลายเป็นโซลูชันหลายองค์ประกอบสำหรับการปฏิบัติตามข้อกำหนดระยะยาว

นอกจากนี้ เอกสารยังแนะนำวิศวกรข้อมูลให้ "เพิ่มประสิทธิภาพการโหลดข้อมูลสำหรับสายข้อมูล" โดย "หลีกเลี่ยงการอัปเดตตาราง BigQuery จำนวนมากด้วยการอัปเดตเล็กๆ น้อยๆ" เนื่องจาก "การอัปเดตแต่ละครั้งจะสร้างเหตุการณ์สายข้อมูล" ซึ่ง "อาจทำให้กราฟสายข้อมูลขยายใหญ่เกินไปและเพิ่มการใช้ DCU และพื้นที่จัดเก็บข้อมูล" และแนะนำให้ใช้คำสั่ง BigQuery MERGE สำหรับการอัปเดตจำนวนมาก [12] นี่แสดงให้เห็นความสัมพันธ์โดยตรงระหว่างรูปแบบการโหลดข้อมูลในไปป์ไลน์ ETL/ELT และค่าใช้จ่ายสายข้อมูลของ Dataplex กลยุทธ์การอัปเดตที่ไม่มีประสิทธิภาพหรือละเอียดอ่อนอาจนำไปสู่การเพิ่มขึ้นของเหตุการณ์สายข้อมูล ซึ่งจะผลักดันค่าใช้จ่ายในการประมวลผล (DCU) และค่าใช้จ่ายในการจัดเก็บเมทาดาตาภายใต้ระดับ Premium Processing วิศวกรข้อมูลจึงต้องออกแบบไปป์ไลน์ข้อมูล BigQuery โดยคำนึงถึงการเพิ่มประสิทธิภาพต้นทุนสายข้อมูล โดยให้ความสำคัญกับการดำเนินงานแบบกลุ่มและคำสั่ง DML ที่มีประสิทธิภาพเพื่อจัดการค่าใช้จ่ายในการดำเนินงานอย่างมีประสิทธิภาพ โดยเฉพาะอย่างยิ่งสำหรับชุดข้อมูลที่มีปริมาณมากและมีการอัปเดตบ่อยครั้ง

อภิธานศัพท์ทางธุรกิจและข้อมูลเชิงลึก

Dataplex มีคุณสมบัติ Business Glossary สำหรับการจัดการคำศัพท์และคำจำกัดความที่เกี่ยวข้องกับธุรกิจ ซึ่งสามารถแนบไปกับคอลัมน์ตารางเพื่อส่งเสริมความเข้าใจที่สอดคล้องกันในการใช้ข้อมูล [6] คุณสมบัตินี้ช่วยปรับปรุงการค้นพบข้อมูลและลดความกำกวม ซึ่งนำไปสู่การวิเคราะห์ที่แม่นยำยิ่งขึ้นและข้อมูลเชิงลึกที่รวดเร็วยิ่งขึ้น [6]

ความสามารถของ Business Glossary ในการจัดการคำศัพท์และคำจำกัดความทางธุรกิจ และการแนบคำศัพท์เหล่านี้เข้ากับคอลัมน์ตาราง [6] ช่วยแก้ไขปัญหาทั่วไป: ความไม่สอดคล้องกันระหว่างโครงสร้างข้อมูลทางเทคนิคและความเข้าใจทางธุรกิจ เมื่อรวมกับ "Data Insights" ที่ขับเคลื่อนด้วย AI และ "Semantic search" [6] จะช่วยให้ผู้ใช้ทางธุรกิจสามารถสำรวจข้อมูลได้อย่างเป็นธรรมชาติมากขึ้น สิ่งนี้บ่งชี้ถึงศักยภาพในการลดเวลาที่วิศวกรข้อมูลใช้ในการแปลคำถามทางธุรกิจเป็นคำสั่งทางเทคนิค หรืออธิบายคำจำกัดความของข้อมูล อย่างไรก็ตาม ยังวางความรับผิดชอบให้วิศวกรข้อมูล

(หรือผู้ดูแลข้อมูลที่พวกเขาร่วมงานด้วย) ต้องมั่นใจว่าสินทรัพย์ข้อมูลทางเทคนิคได้รับการติดแท็กและเชื่อมโยงกับอภิธานศัพท์ทางธุรกิจอย่างถูกต้อง ประสิทธิภาพของข้อมูลเชิงลึกที่ขับเคลื่อนด้วย AI ขึ้นอยู่กับคุณภาพและความสมบูรณ์ของเมทาดาตาบริบททางธุรกิจนี้อย่างมาก ซึ่งส่งเสริมสภาพแวดล้อมการทำงานร่วมกันมากขึ้น โดยวิศวกรข้อมูลมีส่วนร่วมในการทำความเข้าใจข้อมูลนอกเหนือจากการเคลื่อนย้ายและการแปลงข้อมูลเพียงอย่างเดียว

ปรัชญาสถาปัตยกรรมและการนำ Data Mesh ไปใช้

Dataplex ถูกอธิบายว่าเป็น "Data Fabric" อัจฉริยะที่รวมการจัดการข้อมูลในสภาพแวดล้อมที่หลากหลาย (เช่น Data Lake, Data Warehouse, ฐานข้อมูลปฏิบัติการ) เข้าไว้ในเฟรมเวิร์กเดียว [13] แพลตฟอร์มนี้มีบทบาทสำคัญในการนำหลักการสถาปัตยกรรม "Data Mesh" มาปฏิบัติ ซึ่งส่งเสริมการกระจายอำนาจการเป็นเจ้าของข้อมูลโดยเจ้าของข้อมูลโดเมน ในขณะที่ยังคงรักษาการกำกับดูแลแบบรวมศูนย์ [13]

Dataplex ช่วยให้องค์กรสามารถจัดระเบียบข้อมูลอย่างมีเหตุผลเป็น "Lake" (ทำหน้าที่เป็นโดเมน Data Mesh) และ "Zone" (แสดงถึงทีมแต่ละทีมหรือโดเมนย่อย) โดยมี "Assets" ที่แมปกับข้อมูลที่จัดเก็บใน Cloud Storage หรือ BigQuery [15] สถาปัตยกรรมนี้สนับสนุนแนวคิด "Data as a Product" โดยรับรองว่าชุดข้อมูลมีความสะอาด ค้นพบได้ และใช้งานได้ทั่วทั้งโดเมน โดยผู้ผลิตข้อมูลต้องรับผิดชอบต่อคุณภาพและเอกสารประกอบ [13] นอกจากนี้ Dataplex ยังผสานรวมอย่างลึกซึ้งกับ BigLake ซึ่งเป็นเอนจินจัดเก็บ Apache Iceberg ดั้งเดิมของ Google Cloud ซึ่งเป็นรากฐานที่มีการจัดการสำหรับ Lakehouse แบบเปิด [17]

การออกแบบหลักของ Dataplex ที่สอดคล้องกับ Data Mesh โดยเน้น "การจัดการข้อมูลแบบกระจายอำนาจ" พร้อม "การกำกับดูแลแบบรวมศูนย์" [13] และมีโครงสร้างที่ชัดเจน เช่น "Lake" เป็นโดเมนและ "Zone" เป็นโดเมนย่อย [15] ทำให้ Dataplex เป็นแพลตฟอร์มพื้นฐานที่ช่วยให้หลักการเหล่านี้ใช้งานได้จริงภายใน GCP สำหรับองค์กรที่นำสถาปัตยกรรม Data Mesh มาใช้หรือวางแผนที่จะใช้ Dataplex ไม่ใช่แค่เครื่องมือ แต่เป็นแพลตฟอร์มที่ทำให้หลักการเหล่านี้เป็นจริงภายใน GCP สิ่งนี้หมายความว่าวิศวกรข้อมูลที่ทำงานในสภาพแวดล้อมดังกล่าวจะพบว่าโครงสร้างโดยธรรมชาติของ Dataplex เข้ากันได้ดีกับการพัฒนาผลิตภัณฑ์ข้อมูลที่เน้นโดเมน ซึ่งช่วยลดความซับซ้อนในการนำสัญญาข้อมูล การควบคุมการเข้าถึง และการแบ่งปันข้อมูลไปใช้ในทีมที่กระจายอำนาจ ซึ่งอาจช่วยเร่งการนำ Data Mesh ไปใช้และประสบความสำเร็จโดยการให้การสนับสนุนในตัวสำหรับหลักการหลักของมัน

ระบบนิเวศการผสานรวม

Dataplex มีการผสานรวมอย่างลึกซึ้งและอัตโนมัติในการนำเข้าเมทาดาตาจากบริการ Google Cloud ที่หลากหลาย รวมถึง BigQuery, Cloud Storage, Cloud SQL, Spanner, Vertex AI, Pub/Sub, Dataform และ Dataproc Metastore [6]

นอกจากนี้ยังมีความสามารถในการขยายเพื่อรวมเมทาดาตาจากแหล่งข้อมูลบุคคลที่สาม แหล่งข้อมูลภายในองค์กร และแหล่งข้อมูลมัลติคลาวด์ ซึ่งทำได้ผ่าน "Custom Entries" "Custom Entry Types" และ "Aspect Types" ซึ่งกำหนดโครงสร้างสำหรับเมทาดาตาภายนอก [2] สามารถพัฒนาตัวเชื่อมต่อแบบกำหนดเอง (เช่น โดยใช้ PySpark บน Dataproc Serverless) เพื่อดึงเมทาดาตาจากระบบภายนอกและนำเข้าสู่ Dataplex โดยใช้ไฟล์รูปแบบ JSON Lines ผ่าน API

metadataJobs.create [6] นอกจากนี้ยังรองรับการผสานรวม OpenLineage สำหรับการติดตามสายข้อมูลจากระบบภายนอก [6]

แม้ว่า Dataplex จะถูกนำเสนอว่าเป็น "Data Fabric" แบบครบวงจรที่ "ครอบคลุมผู้ให้บริการคลาวด์หลายรายและสภาพแวดล้อมภายในองค์กร" [2] แต่เอกสารรายละเอียดแสดงให้เห็นถึงความแตกต่างที่ชัดเจนในความพยายามในการผสานรวม การดึงเมทาดาตาอัตโนมัติเป็นไปอย่างราบรื่นสำหรับบริการ GCP [6] อย่างไรก็ตาม การผสานรวม "แหล่งข้อมูลบุคคลที่สาม" หรือ "แหล่งข้อมูลแบบกำหนดเอง" ต้องมีการสร้าง "Custom Entries" "Custom Entry Types" และ "Custom Connectors" [6] ซึ่งมักเกี่ยวข้องกับการพัฒนา PySpark และการจัดการไฟล์นำเข้า JSON Lines สิ่งนี้บ่งชี้ว่า Dataplex มีประสิทธิภาพสูงสุดและพร้อมใช้งานทันทีสำหรับองค์กรที่ใช้ GCP เป็นหลัก แม้ว่าการผสานรวมแบบมัลติคลาวด์และไฮบริดจะทำได้ในทางเทคนิค แต่ก็ต้องใช้ต้นทุนรวมในการเป็นเจ้าของ (TCO) ที่สูงขึ้นอย่างมาก เนื่องจากต้องมีการพัฒนา การปรับใช้ และการบำรุงรักษาที่กำหนดเองจากทีมวิศวกรข้อมูลจำนวนมาก สำหรับสภาพแวดล้อมข้อมูลที่มีความหลากหลายอย่างแท้จริง วิศวกรข้อมูลจะต้องชั่งน้ำหนักประโยชน์ของการผสานรวม GCP อย่างลึกซึ้งของ Dataplex กับภาระในการสร้างและบำรุงรักษาไปป์ไลน์การผสานรวมแบบกำหนดเองสำหรับแหล่งข้อมูลที่ไม่ใช่ GCP สิ่งนี้ทำให้ Dataplex เป็นตัวเลือกที่ "เหมาะสมที่สุด" สำหรับองค์กรที่เน้น GCP ซึ่งมีข้อมูลภายนอกบางส่วน มากกว่าจะเป็นโซลูชันการกำกับดูแลแบบมัลติคลาวด์ที่ไม่ขึ้นกับผู้จำหน่ายโดยไม่ต้องลงทุนด้านวิศวกรรมแบบกำหนดเองจำนวนมาก

รูปแบบการกำหนดราคาและการเพิ่มประสิทธิภาพต้นทุน

Dataplex ทำงานบนรูปแบบการเรียกเก็บเงินแบบจ่ายตามการใช้งานจริง (pay-as-you-go) โดยมีปัจจัยหลักในการบริโภคคือการประมวลผลข้อมูล (วัดเป็น Data Compute Unit, DCU, ชั่วโมง) การจัดเก็บเมทาดาตา และการใช้งาน API [12]

- **การประมวลผลข้อมูล (DCU):**

    - **Standard Processing:** ครอบคลุมงานการจัดการข้อมูลพื้นฐาน เช่น การค้นพบและการลงทะเบียนเมทาดาตาจากแหล่งข้อมูล มีการใช้งานฟรี 100 DCU ชั่วโมงแรกต่อเดือน [12]
    - **Premium Processing:** สำหรับความสามารถขั้นสูง เช่น การตรวจสอบคุณภาพข้อมูล การทำโปรไฟล์ข้อมูล และสายข้อมูล ไม่มีระดับฟรีสำหรับ Premium Processing และมีอัตราต่อชั่วโมงต่อ DCU ที่สูงกว่า [12] สายข้อมูลเป็นปัจจัยสำคัญที่ทำให้เกิดค่าใช้จ่าย Premium Processing [12]
- **การจัดเก็บเมทาดาตา:** เมทาดาตาทางเทคนิคที่รวบรวมโดยอัตโนมัติจะถูกจัดเก็บโดยไม่มีค่าใช้จ่าย เมทาดาตาที่เสริม (เช่น Business Glossary, Custom Aspects) จะถูกเรียกเก็บเงิน มีพื้นที่จัดเก็บฟรี 1 MiB แรก จากนั้นคิดค่าบริการ 2 ดอลลาร์ต่อ GiB ต่อเดือน [12]
- **การใช้งาน API:** การโต้ตอบแบบโปรแกรมผ่าน API (Data Catalog API, Data Lineage API) จะถูกเรียกเก็บเงินหลังจากมีการเรียกใช้ฟรี 1 ล้านครั้งแรกต่อเดือน (10 ดอลลาร์ต่อ 100,000 การเรียกใช้) [12]

แบบจำลองการกำหนดราคาแบบแบ่งระดับ (Standard vs. Premium) และปัจจัยที่ทำให้เกิดค่าใช้จ่ายที่ชัดเจนสำหรับการประมวลผล (DCU) การจัดเก็บเมทาดาตา และการใช้งาน API [12] หมายความว่าค่าใช้จ่ายขึ้นอยู่กับ

วิธีการ ใช้คุณสมบัติของ Dataplex อย่างมาก คำแนะนำในการ "เลือกใช้คุณสมบัติอย่างรอบคอบ" "ใช้การสุ่มตัวอย่างและขอบเขตแบบเพิ่มขึ้น" และ "ลดความซับซ้อนของกฎคุณภาพข้อมูล" [12] เชื่อมโยงการตัดสินใจทางวิศวกรรมกับผลลัพธ์ทางการเงินโดยตรง สิ่งนี้บ่งชี้ว่าแนวทาง "เปิดใช้งานทั้งหมด" โดยเฉพาะอย่างยิ่งสำหรับคุณสมบัติ Premium เช่น คุณภาพข้อมูล การทำโปรไฟล์ และสายข้อมูล อาจนำไปสู่ค่าใช้จ่ายที่ไม่คาดคิดและอาจสูง วิศวกรข้อมูลต้องใช้ความคิดที่คำนึงถึงต้นทุนในการออกแบบและการดำเนินงาน ซึ่งหมายถึงการประเมินความจำเป็นและขอบเขตของแต่ละคุณสมบัติอย่างรอบคอบ การควบคุมการรันงานอย่างละเอียด (เช่น การกำหนดเวลา การสุ่มตัวอย่าง) และการเพิ่มประสิทธิภาพรูปแบบการโหลดข้อมูลเพื่อลดการใช้ DCU พวกเขาจำเป็นต้องเข้าใจว่าการเลือกทางเทคนิคของพวกเขาส่งผลโดยตรงต่อรายการการเรียกเก็บเงิน ซึ่งจำเป็นต้องสร้างสมดุลระหว่างการกำกับดูแลที่ครอบคลุมและประสิทธิภาพด้านต้นทุน

กลยุทธ์การเพิ่มประสิทธิภาพต้นทุนสำหรับวิศวกรข้อมูล:
วิศวกรข้อมูลสามารถลดค่าใช้จ่าย Dataplex ได้โดย:

- เลือกใช้คุณสมบัติอย่างรอบคอบ โดยใช้การทำโปรไฟล์และการตรวจสอบคุณภาพเฉพาะในส่วนที่จำเป็นจริงๆ [12]
- ใช้การสุ่มตัวอย่างและขอบเขตแบบเพิ่มขึ้นสำหรับงานทำโปรไฟล์และคุณภาพข้อมูลเพื่อจำกัดปริมาณข้อมูลที่ประมวลผล [12]
- ลดความซับซ้อนของกฎคุณภาพข้อมูล เพื่อลดความซับซ้อนในการประมวลผล [12]
- เพิ่มประสิทธิภาพการโหลดข้อมูลสำหรับสายข้อมูล หลีกเลี่ยงการอัปเดตเล็กๆ น้อยๆ จำนวนมาก และเลือกใช้การดำเนินการแบบกลุ่ม เช่น คำสั่ง BigQuery MERGE [12]
- ตรวจสอบการใช้งานโดยใช้รายงาน Cloud Billing โดยกรองด้วยป้ายกำกับ dataplex เพื่อระบุปัจจัยที่ทำให้เกิดค่าใช้จ่าย เช่น LINEAGE หรือ DATA_PROFILE [12]

จุดแข็งและข้อควรพิจารณาสำหรับวิศวกรข้อมูล

จุดแข็ง:

- **การผสานรวม GCP อย่างลึกซึ้ง:** การนำเข้าเมทาดาตาที่ราบรื่นและอัตโนมัติจากบริการ Google Cloud ที่หลากหลาย ช่วยลดความพยายามในการผสานรวมสำหรับสภาพแวดล้อมข้อมูลที่ใช้ GCP เป็นหลัก [6]
- **การกำกับดูแลที่ขับเคลื่อนด้วย AI:** ใช้ AI/ML สำหรับการค้นพบเมทาดาตา การค้นหาเชิงความหมาย และข้อมูลเชิงลึก ซึ่งอาจทำให้การเข้าถึงข้อมูลเป็นประชาธิปไตยและลดความพยายามในการทำความเข้าใจข้อมูลด้วยตนเอง [16]
- **การสนับสนุน Data Mesh ที่แข็งแกร่ง:** จัดเตรียมโครงสร้างสถาปัตยกรรม (Lake, Zone) ที่สอดคล้องโดยตรงกับหลักการ Data Mesh ซึ่งอำนวยความสะดวกในการกระจายอำนาจการเป็นเจ้าของข้อมูลพร้อมกับการกำกับดูแลแบบรวมศูนย์ [13]
- **การดำเนินงานแบบ Serverless:** ฟังก์ชันการทำงานของ Dataplex จำนวนมากเป็นแบบ Serverless ซึ่งช่วยลดความซับซ้อนของโครงสร้างพื้นฐานพื้นฐานและทำให้การดำเนินงานง่ายขึ้น [2]
- **คุณภาพข้อมูลและสายข้อมูลอัตโนมัติ:** นำเสนอความสามารถที่แข็งแกร่งสำหรับการทำโปรไฟล์ข้อมูล การกำหนดกฎคุณภาพ (ในรูปแบบโค้ด) และการติดตามการไหลของข้อมูลโดยอัตโนมัติสำหรับบริการ GCP [6]
- **Metamodel ที่ขยายได้:** อนุญาตให้ใช้ประเภทรายการและลักษณะที่กำหนดเองเพื่อรวมเมทาดาตาจากแหล่งข้อมูลที่ไม่ใช่ GCP ซึ่งให้ความยืดหยุ่นสำหรับภูมิทัศน์ข้อมูลที่หลากหลาย [6]

ข้อควรพิจารณา/ข้อจำกัด:

- **การเก็บรักษาสายข้อมูลที่จำกัด:** ข้อมูลสายข้อมูลจะถูกเก็บไว้เพียง 30 วันเท่านั้น ซึ่งเป็นข้อจำกัดที่สำคัญสำหรับการปฏิบัติตามข้อกำหนดและการตรวจสอบระยะยาว ซึ่งจำเป็นต้องมีโซลูชันการจัดเก็บภายนอกเพิ่มเติม [20]
- **ความซับซ้อนในการกำหนดราคา:** การกำหนดราคาแบบ DCU แบบแบ่งระดับและปัจจัยที่ทำให้เกิดค่าใช้จ่ายต่างๆ อาจซับซ้อนในการคาดการณ์และจัดการหากไม่มีการตรวจสอบและกลยุทธ์การเพิ่มประสิทธิภาพอย่างรอบคอบ [12]
- **ความพยายามในการผสานรวมแบบมัลติคลาวด์/ไฮบริด:** แม้ว่าจะขยายได้ แต่การรวมแหล่งข้อมูลที่ไม่ใช่ GCP แบบมัลติคลาวด์ หรือแหล่งข้อมูลภายในองค์กรต้องใช้การวิศวกรรมแบบกำหนดเองจำนวนมาก (การสร้างตัวเชื่อมต่อ การจัดการเวิร์กโฟลว์การนำเข้า) ซึ่งเพิ่ม TCO สำหรับสภาพแวดล้อมที่หลากหลาย [2]
- **ขอบเขตภูมิภาค:** ขอบเขตภูมิภาคไม่กว้างขวางเท่ากับคู่แข่งบางราย ซึ่งอาจส่งผลกระทบต่อเวลาแฝงหรือข้อกำหนดด้านความซ้ำซ้อนในบางพื้นที่ทางภูมิศาสตร์ [26]
- **โควตาและข้อจำกัดของระบบ:** มีโควตาคำขอ API และข้อจำกัดของระบบที่เข้มงวดเกี่ยวกับขนาด

และจำนวนรายการ/ลักษณะ ซึ่งจำเป็นต้องพิจารณาสำหรับการใช้งานขนาดใหญ่มาก [27]

แม้ว่า Dataplex จะถูกนำเสนออย่างสม่ำเสมอว่าเป็น "Data Fabric" แบบครบวงจรที่ "ครอบคลุมผู้ให้บริการคลาวด์หลายรายและสภาพแวดล้อมภายในองค์กร" [2] แต่รายละเอียดวิธีการผสานรวมสำหรับแหล่งข้อมูลที่ไม่ใช่ GCP [6] นั้นเกี่ยวข้องกับการสร้าง "ตัวเชื่อมต่อแบบกำหนดเอง" และ "การนำเข้าเมทาดาตาด้วยตนเอง" ผ่าน API ซึ่งแตกต่างอย่างมากจาก "เมทาดาตาที่ดึงมาโดยอัตโนมัติ" สำหรับบริการ GCP ดั้งเดิม [6] ความไม่สอดคล้องกันนี้เน้นย้ำว่าแม้ Dataplex

สามารถ ผสานรวมแหล่งข้อมูลภายนอกได้ แต่ความพยายามที่ต้องใช้ก็สูงกว่ามากสำหรับข้อมูลที่ใช้ GCP เป็นหลัก สำหรับวิศวกรข้อมูล สิ่งนี้หมายความว่าวิสัยทัศน์ "Data Fabric" แบบครบวงจรนั้นสามารถทำได้ง่ายกว่าและคุ้มค่ากว่าในสภาพแวดล้อมที่ใช้ GCP เป็นหลัก การขยายไปยังคลาวด์อื่นหรือระบบภายในองค์กรต้องใช้ค่าใช้จ่ายในการพัฒนาและบำรุงรักษาที่สูงมาก ทำให้เป็นแพลตฟอร์มที่เน้น GCP เป็นอันดับแรกพร้อมความสามารถในการขยาย มากกว่าจะเป็นโซลูชันมัลติคลาวด์ที่ไม่ขึ้นกับผู้จำหน่ายอย่างแท้จริงโดยไม่ต้องลงทุนด้านวิศวกรรมแบบกำหนดเองจำนวนมาก นี่เป็นข้อควรพิจารณาที่สำคัญสำหรับองค์กรที่มีกลยุทธ์มัลติคลาวด์เป็นอันดับแรกอย่างแท้จริง

Microsoft Purview: มุมมองของวิศวกรข้อมูล

Microsoft Purview เป็นชุดผลิตภัณฑ์ที่ครอบคลุมซึ่งครอบคลุมโซลูชันการกำกับดูแลข้อมูล ความปลอดภัยของข้อมูล และความเสี่ยงและการปฏิบัติตามข้อกำหนด [1]

ความสามารถหลักและคุณสมบัติ

แค็ตตาล็อกแบบรวมศูนย์และ Data Map

Microsoft Purview เป็นพอร์ตโฟลิโอที่ครอบคลุมสำหรับการกำกับดูแลข้อมูล ความปลอดภัยของข้อมูล และความเสี่ยงและการปฏิบัติตามข้อกำหนด [1] Unified Catalog และ Data Map เป็นหัวใจสำคัญของโซลูชันการกำกับดูแลข้อมูล โดยมอบประสบการณ์ที่ทันสมัยเพื่อการมองเห็นที่ครอบคลุมและความน่าเชื่อถือของข้อมูล [7]

Data Map เป็นองค์ประกอบพื้นฐานสำหรับการค้นพบข้อมูลและการกำกับดูแล โดยรวบรวมเมทาดา

ตาจากระบบการวิเคราะห์, SaaS และระบบปฏิบัติการในสภาพแวดล้อมแบบไฮบริด, ภายในองค์กร และมัลติคลาวด์ [29] Data Map รักษาความทันสมัยผ่านระบบการสแกนและการจัดประเภทแบบรวมศูนย์ ซึ่งสนับสนุนการค้นพบและการจัดประเภทข้อมูลโดยอัตโนมัติ [8] Unified Catalog ช่วยให้สามารถสร้างโดเมนการกำกับดูแล จัดการผลิตภัณฑ์ข้อมูล และเชื่อมโยงข้อมูลกับแนวคิดทางธุรกิจ (OKRs, คำศัพท์ในอภิธานศัพท์) [7] Data Map ปรับขนาดได้อย่างยืดหยุ่นตามความต้องการ โดยมีหน่วยความจุ (CU) สำหรับปริมาณงาน (25 การดำเนินการ/วินาทีต่อ CU) และการจัดเก็บเมทาดาตา (10 GB ต่อ CU) [29]

การนำเสนอ Purview ในฐานะ "พอร์ตโฟลิโอที่ครอบคลุม" [1] ที่รวม "โซลูชันความปลอดภัยข้อมูล และความเสี่ยงและการปฏิบัติตามข้อกำหนด" เข้ากับการกำกับดูแลข้อมูลอย่างชัดเจน ถือเป็นจุดเด่นที่สำคัญ คุณสมบัติที่รวมเข้าด้วยกัน เช่น Data Loss Prevention (DLP), Information Protection และ Insider Risk Management [3] แสดงให้เห็นถึงแนวทางแบบองค์รวม สิ่งนี้หมายความว่าวิศวกรข้อมูลสามารถใช้แพลตฟอร์มเดียวสำหรับการค้นพบข้อมูลและสายข้อมูล รวมถึงการบังคับใช้นโยบายข้อมูลที่ละเอียดอ่อนและการควบคุมการปฏิบัติตามข้อกำหนด ซึ่งช่วยลดความซับซ้อนในการจัดการเครื่องมือรักษาความปลอดภัยที่แตกต่างกัน และอาจลดความจำเป็นในการผสานรวมแบบกำหนดเองสำหรับการปกป้องข้อมูล สำหรับองค์กรในอุตสาหกรรมที่มีการควบคุมอย่างเข้มงวด ชุดโซลูชันแบบรวมศูนย์นี้ช่วยลดความพยายามทางวิศวกรรมที่จำเป็นในการปฏิบัติตามข้อกำหนดด้านความปลอดภัยและการปฏิบัติตามข้อกำหนดที่ซับซ้อน โดยมอบระนาบควบคุมแบบรวมศูนย์สำหรับการจัดการความเสี่ยงข้อมูล

ความสามารถในการ "ปรับขนาดได้อย่างยืดหยุ่น" และ "ปรับขนาดอัตโนมัติ" ของ Data Map โดยมีการเรียกเก็บเงินตามหน่วยความจุ (CU) สำหรับการดำเนินการและพื้นที่จัดเก็บ [29] แสดงให้เห็นถึงสถาปัตยกรรมที่ยืดหยุ่นซึ่งออกแบบมาสำหรับความต้องการในการประมวลผลและจัดเก็บเมทาดาตาที่ผันผวน การออกแบบนี้หมายความว่าวิศวกรข้อมูลอาจมีภาระการดำเนินงานที่เกี่ยวข้องกับการวางแผนความจุสำหรับที่เก็บเมทาดาตาน้อยลง Purview สามารถปรับให้เข้ากับการใช้งานที่เพิ่มขึ้นอย่างรวดเร็ว (เช่น การสแกนเริ่มต้นจำนวนมาก การไหลเข้าของเมทาดาตาใหม่ทันที) ซึ่งอาจรับประกันประสิทธิภาพที่สอดคล้องกันโดยไม่ต้องมีการแทรกแซงด้วยตนเอง อย่างไรก็ตาม การกล่าวถึงการต้อง "ขอโควตา" เพื่อ "เพิ่มหน้าต่างความยืดหยุ่น" สำหรับปริมาณงานที่สูงขึ้น [29] ชี้ให้เห็นว่าแม้การปรับขนาดส่วนใหญ่จะเป็นไปโดยอัตโนมัติ แต่สถานการณ์ที่มีปริมาณงานสูงมากหรือต่อเนื่องอาจยังคงต้องมีการคาดการณ์และการแทรกแซงจากผู้ดูแลระบบ นี่เป็นข้อควรพิจารณาที่สำคัญสำหรับสถาปนิกที่วางแผนสำหรับสภาพแวดล้อมข้อมูลองค์กรขนาดใหญ่

การตรวจสอบคุณภาพข้อมูลและข้อมูลเชิงลึกอัตโนมัติ

Purview นำเสนอความสามารถด้านคุณภาพข้อมูลในตัว ซึ่งช่วยให้เจ้าของข้อมูลสามารถกำกับดูแลและปรับปรุงคุณภาพของระบบนิเวศข้อมูลของตนได้ [7] มีข้อมูลเชิงลึกด้านคุณภาพข้อมูลอัตโนมัติและคุณสมบัติในการตรวจสอบและแก้ไขบันทึกข้อผิดพลาดด้านคุณภาพข้อมูล รวมถึงการจัดการข้อยกเว้น

ของกฎ [8] คุณสมบัติการสังเกตการณ์ข้อมูลอยู่ในช่วงพรีวิวสำหรับผู้ดูแลข้อมูลและเจ้าของผลิตภัณฑ์ข้อมูล [33] การตรวจจับรูปแบบอัตโนมัติ (Parquet, Delta, Iceberg) มีให้บริการทั่วไปสำหรับการสแกนคุณภาพข้อมูล ซึ่งช่วยลดความซับซ้อนในการตั้งค่าสำหรับผู้ดูแลคุณภาพข้อมูล [33] การจัดการสุขภาพข้อมูล รวมถึงคุณภาพข้อมูล จะถูกเรียกเก็บเงินตามหน่วยประมวลผลการกำกับดูแลข้อมูล (DGPU) [34]

เอกสารของ Purview เน้นย้ำถึง "ประสบการณ์คุณภาพข้อมูลในตัวเพื่อเสริมสร้างศักยภาพให้เจ้าของข้อมูลกำกับดูแลระบบนิเวศข้อมูลของตน" [7] และช่วยให้ "วิศวกรข้อมูล ผู้ดูแลคุณภาพข้อมูล และนักวิเคราะห์สามารถตรวจสอบและแก้ไขข้อมูล รวมถึงติดตามการปรับปรุงอย่างต่อเนื่อง" [33] สิ่งนี้บ่งชี้ถึงรูปแบบการจัดการคุณภาพข้อมูลแบบรวมศูนย์หรือกระจายอำนาจมากขึ้น โดยที่เจ้าของข้อมูลทางธุรกิจมีอินเทอร์เฟซโดยตรงสำหรับการโต้ตอบ แนวทางนี้หมายความว่าวิศวกรข้อมูลอาจมีส่วนร่วมน้อยลงในการกำหนดและแก้ไขปัญหาคุณภาพข้อมูลประจำวัน เนื่องจากความรับผิดชอบบางส่วนถูกโอนไปยังเจ้าของข้อมูลทางธุรกิจหรือผู้ดูแลข้อมูลโดยเฉพาะ ในขณะที่วิศวกรยังคงรับผิดชอบในการสร้างไปป์ไลน์ข้อมูลพื้นฐานและรับรองความแข็งแกร่งของกรอบการทำงานคุณภาพข้อมูล บทบาทของพวกเขาอาจพัฒนาไปสู่การเป็นผู้สนับสนุนและแก้ไขปัญหาสำหรับกระบวนการคุณภาพข้อมูลที่จัดการโดยผู้มีส่วนได้ส่วนเสียที่ไม่ใช่ด้านเทคนิค ซึ่งสามารถเพิ่มเวลาให้วิศวกรทำงานที่ซับซ้อนมากขึ้นได้ แต่ต้องมีไปป์ไลน์ที่สามารถผสานรวมและตอบสนองต่อข้อเสนอแนะด้านคุณภาพที่ขับเคลื่อนโดยเจ้าของข้อมูล

สายข้อมูล (Data Lineage)

Purview ให้บริการสายข้อมูลอัตโนมัติสำหรับบริการ Azure จำนวนมาก รวมถึง Azure Data Factory, Azure Data Share, Azure Synapse, Power BI และ Microsoft Fabric [9] สายข้อมูลช่วยระบุความสัมพันธ์ระหว่างผลิตภัณฑ์ข้อมูลและติดตามสาเหตุหลักของปัญหาคุณภาพ [7]

การรายงานสายข้อมูลแบบกำหนดเองได้รับการสนับสนุนผ่าน Apache Atlas API hooks และ REST API สำหรับสถานการณ์ที่สายข้อมูลอัตโนมัติไม่สมบูรณ์หรือขาดหายไป [35] รองรับทั้งสายข้อมูลระดับสินทรัพย์และระดับคอลัมน์ โดยมีตัวเลือกสายข้อมูลด้วยตนเองสำหรับแหล่งที่มาที่ยังไม่รองรับระบบอัตโนมัติ [36] แคนวาสสายข้อมูลมีการแสดงภาพ โดยมีมุมมองเริ่มต้นห้าระดับที่สามารถขยายได้ [36]

แม้ว่า Purview จะนำเสนอสายข้อมูลอัตโนมัติสำหรับบริการ Azure ดั้งเดิม [35] แต่ก็ยอมรับอย่างชัดเจนถึงสถานการณ์ที่ "สายข้อมูลที่สร้างขึ้นโดย Purview โดยอัตโนมัติไม่สมบูรณ์หรือขาดหายไป" และมี "สายข้อมูลด้วยตนเอง" หรือ "Apache Atlas API/REST API" สำหรับการรายงานแบบกำหนดเอง [36] สิ่งนี้สะท้อนถึงแนวทางของ Dataplex สำหรับแหล่งข้อมูลภายนอก ซึ่งบ่งชี้ว่าการบรรลุสายข้อมูลแบบครบวงจรอย่างแท้จริงในสภาพแวดล้อมข้อมูลที่ซับซ้อนสูงหรือหลากหลาย (โดยเฉพาะอย่างยิ่งที่เกี่ยวข้องกับการแปลงแบบกำหนดเองที่ไม่ใช่ Azure หรือเครื่องมือของบุคคลที่สาม) ยังคงต้องใช้ความพยายามในการพัฒนาอย่างมีนัยสำคัญจากวิศวกรข้อมูล พวกเขาจะต้องใช้ Apache Atlas

API เพื่อส่งเมทาดาตาของสายข้อมูลสำหรับกระบวนการ ETL/ELT แบบกำหนดเองหรือระบบภายนอก ซึ่งเพิ่มภาระในการพัฒนาและบำรุงรักษา ซึ่งหมายความว่าการอ้างสิทธิ์ "แบบครบวงจร" นั้นเป็นจริงสำหรับบริการ Azure ดั้งเดิมเป็นหลัก และวิศวกรจะต้องเติมเต็มช่องว่างอย่างแข็งขันเพื่อภาพรวมที่สมบูรณ์ทั่วทั้งภูมิทัศน์ข้อมูลของตน

ความปลอดภัยของข้อมูลและการปฏิบัติตามข้อกำหนด

Purview เป็นชุดโซลูชันที่ครอบคลุมซึ่งครอบคลุมความปลอดภัยของข้อมูล ความเสี่ยง และการปฏิบัติตามข้อกำหนด โดยให้ความคุ้มครองแบบรวมศูนย์ [1] คุณสมบัติความปลอดภัยที่สำคัญ ได้แก่ Data Loss Prevention (DLP), Information Protection, Insider Risk Management และ Data Security Posture Management (DSPM) สำหรับ AI [3]

แพลตฟอร์มนี้ให้การเข้าถึงแบบโปรแกรมไปยังเอนจินการประเมินนโยบายเพื่อการบังคับใช้นโยบายความปลอดภัยข้อมูลและการกำกับดูแลที่สอดคล้องกัน [31] รองรับการจัดประเภทอัตโนมัติและการกำหนดป้ายกำกับความละเอียดอ่อน รวมถึงการจัดการการปฏิบัติตามข้อกำหนดสำหรับกฎระเบียบต่างๆ เช่น GDPR, HIPAA, SOX และ ISO 27001 [3]

การรวม DLP, Information Protection และ Insider Risk Management เข้ากับแพลตฟอร์มการกำกับดูแลอย่างชัดเจนและลึกซึ้ง [3] ถือเป็นข้อได้เปรียบที่สำคัญเหนือ Dataplex ซึ่งมุ่งเน้นไปที่ Data Fabric และการกำกับดูแลที่แคบกว่า แนวทางแบบรวมศูนย์นี้หมายความว่าวิศวกรข้อมูลสามารถใช้แพลตฟอร์มเดียวสำหรับการค้นพบข้อมูล การจัดประเภท และการบังคับใช้นโยบายข้อมูลที่ละเอียดอ่อน ซึ่งช่วยลดความซับซ้อนในการจัดการเครื่องมือรักษาความปลอดภัยที่แตกต่างกัน และปรับปรุงความพยายามในการปฏิบัติตามข้อกำหนดโดยให้มุมมองที่เป็นหนึ่งเดียวของความเสี่ยงข้อมูลและการบังคับใช้นโยบายอัตโนมัติ สำหรับองค์กรที่มีข้อกำหนดด้านกฎระเบียบที่เข้มงวดหรือเน้นการปกป้องข้อมูลเชิงรุก Purview นำเสนอโซลูชันที่ครอบคลุมและพร้อมใช้งานทันที ซึ่งช่วยลดความจำเป็นในการผสานรวมความปลอดภัยแบบกำหนดเองในไปป์ไลน์ข้อมูลของตน

อย่างไรก็ตาม Purview มี "ข้อจำกัด" เฉพาะของระบบเกี่ยวกับ "จำนวนสูงสุดของ SITs แบบกำหนดเอง" "จำนวนสูงสุดของนโยบายต่อผู้เช่า" และ "จำนวนสูงสุดของป้ายกำกับการเก็บรักษา" [40] ตัวอย่างเช่น ผู้เช่ารายเดียวมีนโยบายสูงสุด 10,000 นโยบายในทุกประเภทการปฏิบัติตามข้อกำหนด สำหรับองค์กรขนาดใหญ่หรือซับซ้อนมากที่มีแหล่งข้อมูลหลายพันแห่ง การควบคุมการเข้าถึงที่ละเอียดอ่อนหรือข้อกำหนดด้านกฎระเบียบที่กว้างขวาง ข้อจำกัดเหล่านี้อาจกลายเป็นข้อจำกัดได้ วิศวกรข้อมูลและสถาปนิกจะต้องวางแผนการกำหนดนโยบายภายใน Purview อย่างรอบคอบเพื่อหลีกเลี่ยงการชนกับขีดจำกัดเหล่านี้ ซึ่งอาจจำเป็นต้องมีแนวทางเชิงกลยุทธ์ในการรวมนโยบาย การจัดกลุ่มกฎ หรือการจัดการข้อยกเว้น ซึ่งอาจส่งผลกระทบต่อระดับความละเอียดที่ต้องการสำหรับสถานการณ์การปฏิบัติตามข้อกำหนดบางอย่าง วิศวกรจำเป็นต้องตระหนักถึงข้อจำกัดเหล่านี้ในระหว่างการออกแบบโซลูชัน

และการปรับขนาด

ปรัชญาสถาปัตยกรรมและ Data as a Product

Purview ส่งเสริม "แนวทางการกำกับดูแลแบบรวมศูนย์" โดยสร้างสมดุลระหว่างการกำกับดูแลแบบรวมศูนย์ (สำหรับกฎ ความปลอดภัย คุณภาพ และมาตรฐาน) กับความรับผิดชอบแบบกระจายอำนาจและความสามารถในการบริการตนเองสำหรับการเข้าถึงข้อมูล การค้นพบ และการบำรุงรักษา [7] แพลตฟอร์มนี้เน้นย้ำอย่างมากถึงโมเดล "Data as a Product" (DaaP) ซึ่งข้อมูลถูกดูแล กำกับดูแล และส่งมอบด้วยความตั้งใจที่ชัดเจนเพื่อตอบสนองวัตถุประสงค์ทางธุรกิจที่เฉพาะเจาะจง เป็นไปตามมาตรฐานคุณภาพ และสามารถนำกลับมาใช้ใหม่ได้ทั่วทั้งทีม [9] โมเดลนี้กำหนดความเป็นเจ้าของที่ชัดเจน เมทาดาตาที่สอดคล้องกับธุรกิจ กฎคุณภาพ มาตรการการปฏิบัติตามข้อกำหนด และเวิร์กโฟลว์การเข้าถึงที่กำหนดไว้สำหรับผลิตภัณฑ์ข้อมูลแต่ละรายการ [11] Unified Catalog ของ Purview รองรับ "โดเมนการกำกับดูแล" "ผลิตภัณฑ์ข้อมูล" "องค์ประกอบข้อมูลที่สำคัญ" "คำศัพท์ในอภิธานศัพท์" และ "วัตถุประสงค์และผลลัพธ์หลัก (OKRs)" เพื่อเชื่อมโยงข้อมูลกับเป้าหมายทางธุรกิจ [9] Purview เป็นโซลูชัน Platform-as-a-Service (PaaS) ที่มีจุดเชื่อมต่อสาธารณะที่สามารถเข้าถึงได้ผ่านอินเทอร์เน็ต แม้ว่าจะมีตัวเลือกการเชื่อมต่อส่วนตัวก็ตาม [44]

การที่ Purview สอดคล้องอย่างแน่นแฟ้นกับโมเดล "Data as a Product" (DaaP) [9] แสดงให้เห็นว่าแพลตฟอร์มนี้ได้รับการออกแบบมาเพื่อเปิดใช้งานการเปลี่ยนแปลงองค์กรและวัฒนธรรมเฉพาะในการจัดการและการบริโภคข้อมูล โมเดลนี้เน้นย้ำว่าผลลัพธ์ข้อมูลควรได้รับการปฏิบัติอย่างเคร่งครัดเช่นเดียวกับผลิตภัณฑ์ซอฟต์แวร์ โดยมีเจ้าของที่ชัดเจน กฎคุณภาพ และการเข้าถึงที่กำหนดไว้ การเปลี่ยนแปลงนี้บ่งบอกถึงวิวัฒนาการที่สำคัญในบทบาทของวิศวกรข้อมูล พวกเขาไม่ได้เพียงแค่สร้างไปป์ไลน์เพื่อย้ายและแปลงข้อมูลเท่านั้น แต่ยังร่วมสร้าง "ผลิตภัณฑ์ที่สร้างขึ้นตามวัตถุประสงค์" [11] ซึ่งต้องสามารถค้นพบได้ มีคุณภาพสูง และเป็นไปตามข้อกำหนด การเปลี่ยนแปลงนี้จำเป็นต้องมีการทำงานร่วมกันอย่างใกล้ชิดกับ "เจ้าของผลิตภัณฑ์ข้อมูล" และ "ผู้ดูแลข้อมูล" [11] เพื่อกำหนดเมทาดาตาที่สอดคล้องกับธุรกิจ มาตรฐานคุณภาพ และเวิร์กโฟลว์การเข้าถึง ซึ่งส่งเสริมให้วิศวกรใช้ความคิดที่เน้นผลิตภัณฑ์ โดยมุ่งเน้นที่ความสามารถในการบริโภคและคุณค่าที่แท้จริงของผลลัพธ์ข้อมูลของตน ซึ่งนำไปสู่การสอดคล้องกับความต้องการทางธุรกิจที่ดีขึ้นและความน่าเชื่อถือของข้อมูลที่เพิ่มขึ้น

ระบบนิเวศการผสานรวม

Purview มีการผสานรวมอย่างแน่นแฟ้นกับบริการ Azure ที่หลากหลาย รวมถึง Azure Data Factory, Azure Synapse Analytics, Azure SQL Database, Azure Data Lake Storage

Gen2 และ Power BI นอกจากนี้ยังผสานรวมกับ Microsoft Fabric ด้วย [7]

Data Map รองรับการสแกนอัตโนมัติของแหล่งข้อมูลภายในองค์กร (เช่น SQL Server, Oracle, File Share ผ่าน Self-Hosted Integration Runtimes) และสภาพแวดล้อมมัลติคลาวด์ (เช่น AWS S3, Google BigQuery) [3] สามารถสร้างตัวเชื่อมต่อแบบกำหนดเอง (เช่น ใน Power Automate/Apps) เพื่อผสานรวมกับระบบภายนอก และมี REST APIs (อิงตาม Apache Atlas) สำหรับสายข้อมูลแบบกำหนดเองและการจัดการเมทาดาตา [35] นอกจากนี้ยังรองรับการเชื่อมต่อที่ปลอดภัยผ่าน Private Endpoints สำหรับทั้งบัญชี Purview และการนำเข้าแหล่งข้อมูล เพื่อให้มั่นใจว่าข้อมูลยังคงอยู่ในเครือข่ายส่วนตัว [44]

จุดแข็งของ Purview อยู่ที่การผสานรวมอัตโนมัติอย่างลึกซึ้งภายในระบบนิเวศของ Azure [7] อย่างไรก็ตาม ยังรองรับ "การสแกนอัตโนมัติของแหล่งข้อมูลภายในองค์กร มัลติคลาวด์ (AWS S3, Google BigQuery) และ SaaS" อย่างชัดเจน [30] แม้ว่าจะให้ความครอบคลุมเมทาดาตาที่กว้างขวาง แต่ความลึกของความสามารถในการกำกับดูแลอัตโนมัติ (เช่น สายข้อมูลหรือการตรวจสอบคุณภาพ) อาจยังคงแข็งแกร่งที่สุดสำหรับบริการ Azure ดั้งเดิม โดยมักจะต้องมีการพัฒนา API แบบกำหนดเองสำหรับการผสานรวมการไหลของข้อมูลที่ไม่ใช่ Azure อย่างสมบูรณ์ [36] สิ่งนี้บ่งชี้ว่า Purview สามารถนำเสนอเส้นทางที่รวดเร็วกว่าในการสร้าง

มุมมองเมทาดาตาแบบรวมศูนย์ ทั่วทั้งสภาพแวดล้อมข้อมูลที่มีความหลากหลายอย่างแท้จริง อย่างไรก็ตาม วิศวกรข้อมูลจะต้องประเมินอย่างรอบคอบว่า "การสแกน" นี้ให้ความลึกของการกำกับดูแลที่เพียงพอหรือไม่ (เช่น สายข้อมูลระดับคอลัมน์อัตโนมัติหรือการบังคับใช้คุณภาพข้อมูลที่ละเอียดอ่อน) สำหรับสินทรัพย์ข้อมูลที่สำคัญที่อยู่ภายนอก Azure หากต้องการการกำกับดูแลอัตโนมัติอย่างลึกซึ้งสำหรับแหล่งข้อมูลที่ไม่ใช่ Azure วิศวกรควรคาดการณ์ความพยายามในการพัฒนาแบบกำหนดเองที่คล้ายคลึงกัน (เช่น การใช้ Apache Atlas APIs) เช่นเดียวกับ Dataplex แม้ว่า Purview จะอ้างว่ามีการสแกนเริ่มต้นที่กว้างกว่าก็ตาม ซึ่งหมายถึงการทำความเข้าใจการแลกเปลี่ยนระหว่างการจัดทำแค็ตตาล็อกที่กว้างขวางและการกำกับดูแลอัตโนมัติอย่างลึกซึ้งทั่วทั้งแพลตฟอร์มที่หลากหลาย

รูปแบบการกำหนดราคาและการเพิ่มประสิทธิภาพต้นทุน

Purview ทำงานบนรูปแบบการเรียกเก็บเงินแบบจ่ายตามการใช้งานจริง (pay-as-you-go) [34]

- **การเรียกเก็บเงิน Unified Catalog:** คิดค่าบริการสำหรับ "สินทรัพย์ที่อยู่ภายใต้การกำกับดูแล" ซึ่งเป็นสินทรัพย์ข้อมูลที่ไม่ซ้ำกันที่เชื่อมโยงกับแนวคิดการกำกับดูแล เช่น ผลิตภัณฑ์ข้อมูล หรือองค์ประกอบข้อมูลที่สำคัญ สินทรัพย์ที่รวบรวมใน Data Map แต่ไม่ได้เชื่อมโยงกับแนวคิดการกำกับดูแลจะไม่ถูกเรียกเก็บเงิน [34]
- **การจัดการสุขภาพข้อมูล:** รวมถึงคุณภาพข้อมูล จะถูกเรียกเก็บเงินตามหน่วยประมวลผลการกำกับดูแลข้อมูล (DGPU) [34] DGPU เป็นหน่วยประมวลผลที่มีการจัดการเต็มรูปแบบที่ใช้ในการรันความสามารถที่ต้องใช้การประมวลผลหนัก เช่น คุณภาพข้อมูลและการจัดการสุขภาพข้อมูล

DGPU ที่ใช้ขึ้นอยู่กับประเภทกฎ (สำเร็จรูปหรือกำหนดเอง) ปริมาณข้อมูล และประเภทแหล่งที่มา [34]

- **Data Map และค่าใช้จ่ายในการสแกน:** ไม่มีการเรียกเก็บเงินสำหรับการสแกนสินทรัพย์เข้าสู่ Data Map หากเป็นไปตามเงื่อนไขการเรียกเก็บเงินบางประการ [34] Data Map มีหน่วยความจุ (CU) สำหรับปริมาณงาน (25 การดำเนินการ/วินาทีต่อ CU) และการจัดเก็บเมทาดาตา (10 GB ต่อ CU) [29]
- **การทดลองใช้ฟรี:** มีเวอร์ชันฟรีของโซลูชันการกำกับดูแลของ Purview ที่มีข้อจำกัด เช่น สินทรัพย์ที่ติดคำอธิบายประกอบสูงสุด 1,000 รายการ และไม่สามารถเปิดตั๋วสนับสนุนได้ [29]

การเรียกเก็บเงินของ Purview สำหรับคุณภาพข้อมูลและการจัดการสุขภาพข้อมูลอิงตาม Data Governance Processing Units (DGPU) ซึ่งได้รับอิทธิพลจากประเภทกฎ ปริมาณข้อมูล และประเภทแหล่งที่มา [34] สิ่งนี้หมายความว่าเช่นเดียวกับ Dataplex วิศวกรข้อมูลต้องคำนึงถึงต้นทุนการคำนวณของการตรวจสอบคุณภาพข้อมูลและสุขภาพข้อมูล การออกแบบกฎที่มีประสิทธิภาพ การใช้การสุ่มตัวอย่าง และการทำความเข้าใจผลกระทบของปริมาณข้อมูลต่อการใช้ DGPU เป็นสิ่งสำคัญสำหรับการจัดการต้นทุน ความสามารถในการ "จัดการตารางเวลาได้อย่างเต็มที่และปิดการควบคุมบางอย่างเพื่อเพิ่มประสิทธิภาพต้นทุน" [34] ช่วยให้วิศวกรมีเครื่องมือในการควบคุมค่าใช้จ่ายในการดำเนินงาน

แบบจำลองการเรียกเก็บเงินสำหรับ "สินทรัพย์ที่อยู่ภายใต้การกำกับดูแล" ซึ่งกำหนดเป็นสินทรัพย์ข้อมูลที่ไม่ซ้ำกันที่ "เชื่อมโยงกับแนวคิดการกำกับดูแล เช่น ผลิตภัณฑ์ข้อมูลหรือองค์ประกอบข้อมูลที่สำคัญ" โดยสินทรัพย์ที่รวบรวมใน Data Map แต่ *ไม่ได้* เชื่อมโยงจะไม่ถูกเรียกเก็บเงิน [34] แบบจำลองการเรียกเก็บเงินนี้กระตุ้นให้มีการจัดการแค็ตตาล็อกอย่างมีกลยุทธ์ วิศวกรข้อมูลควรทำงานร่วมกับเจ้าของข้อมูลและผู้ดูแลข้อมูลเพื่อจัดลำดับความสำคัญของสินทรัพย์ที่ต้องการการกำกับดูแลอย่างเต็มที่ (เช่น การเชื่อมโยงกับแนวคิดทางธุรกิจ) เพื่อจัดการต้นทุนอย่างมีประสิทธิภาพ สิ่งนี้ไม่สนับสนุนแนวทาง "แค็ตตาล็อกทุกอย่าง" หากไม่ใช่ทุกสินทรัพย์ที่ต้องการการกำกับดูแลอย่างกระตือรือร้น ซึ่งผลักดันให้วิศวกรมุ่งเน้นความพยายามไปที่ผลิตภัณฑ์ข้อมูลที่มีมูลค่าสูงและองค์ประกอบข้อมูลที่สำคัญ ซึ่งจำเป็นต้องมีความเข้าใจอย่างลึกซึ้งเกี่ยวกับลำดับความสำคัญทางธุรกิจเพื่อเพิ่มประสิทธิภาพทั้งประสิทธิภาพการกำกับดูแลและประสิทธิภาพด้านต้นทุน

จุดแข็งและข้อควรพิจารณาสำหรับวิศวกรข้อมูล

จุดแข็ง:

- **การกำกับดูแลที่ครอบคลุม:** นำเสนอชุดโซลูชันที่ครบวงจรสำหรับการกำกับดูแลข้อมูล ความปลอดภัย และการปฏิบัติตามข้อกำหนด ซึ่งช่วยลดความซับซ้อนในการจัดการเครื่องมือที่แตกต่างกัน [1]
- **การผสานรวม Azure ที่แข็งแกร่ง:** การผสานรวมอย่างลึกซึ้งและอัตโนมัติกับบริการ Azure ที่

หลากหลาย ช่วยลดความพยายามในการผสานรวมสำหรับสภาพแวดล้อมข้อมูลที่ใช้ Azure เป็นหลัก [7]

- **แนวทางการกำกับดูแลแบบรวมศูนย์ (Federated Governance):** สนับสนุนรูปแบบ Data as a Product ที่ส่งเสริมการกระจายอำนาจความรับผิดชอบข้อมูลพร้อมกับการควบคุมแบบรวมศูนย์ [7]
- **ข้อมูลเชิงลึกที่ขับเคลื่อนด้วย AI:** ใช้ AI/ML เพื่อการจัดประเภทข้อมูลอัตโนมัติ การตรวจสอบคุณภาพ และการระบุความเสี่ยง [8]
- **รองรับสภาพแวดล้อมที่หลากหลาย:** Data Map สามารถสแกนและรวบรวมเมทาดาตาจากแหล่งข้อมูลภายในองค์กร มัลติคลาวด์ (AWS S3, Google BigQuery) และ SaaS ได้ [29]
- **การเชื่อมต่อที่ปลอดภัย:** รองรับ Private Endpoints สำหรับการเข้าถึงบัญชี Purview และการนำเข้าข้อมูล ทำให้มั่นใจได้ถึงความปลอดภัยของเครือข่าย [44]

ข้อควรพิจารณา/ข้อจำกัด:

- **เส้นทางการเรียนรู้:** อาจมีเส้นทางการเรียนรู้ที่สูงชันสำหรับผู้ใช้ใหม่ โดยเฉพาะอย่างยิ่งเนื่องจากความกว้างของชุดคุณสมบัติ [50]
- **การพึ่งพาผู้จำหน่าย:** แม้จะรองรับมัลติคลาวด์ แต่การผสานรวมอย่างลึกซึ้งที่สุดและการทำงานอัตโนมัติยังคงอยู่ในระบบนิเวศของ Microsoft ซึ่งอาจนำไปสู่การพึ่งพาผู้จำหน่ายสำหรับองค์กรที่ใช้เครื่องมือที่ไม่ใช่ Microsoft อย่างมาก [50]
- **ข้อจำกัดของนโยบาย:** มีข้อจำกัดเฉพาะของระบบเกี่ยวกับจำนวนสูงสุดของนโยบาย ป้ายกำกับการเก็บรักษา และ SITs (Sensitive Information Types) ซึ่งอาจเป็นข้อจำกัดสำหรับองค์กรขนาดใหญ่ที่มีข้อกำหนดด้านการกำกับดูแลที่ละเอียดอ่อนมาก [40]
- **ปัญหาประสิทธิภาพ:** มีรายงานปัญหาประสิทธิภาพที่ช้าในบางกรณี ซึ่งอาจส่งผลกระทบต่อประสิทธิภาพการทำงาน [50]
- **ความพยายามในการผสานรวมแบบกำหนดเอง:** แม้จะสามารถสแกนแหล่งข้อมูลที่หลากหลายได้ แต่การบรรลุการกำกับดูแลอัตโนมัติอย่างลึกซึ้ง (เช่น สายข้อมูลระดับคอลัมน์หรือการบังคับใช้คุณภาพข้อมูลแบบละเอียด) สำหรับแหล่งข้อมูลที่ไม่ใช่ Azure ยังคงต้องใช้ความพยายามในการพัฒนาแบบกำหนดเองอย่างมีนัยสำคัญ [36]

แม้ว่า Purview จะนำเสนอความสามารถในการ "สแกนอัตโนมัติของแหล่งข้อมูลภายในองค์กร มัลติคลาวด์ (AWS S3, Google BigQuery) และ SaaS" [30] ซึ่งให้ความครอบคลุมเมทาดาตาที่กว้างขวาง แต่ความลึกของความสามารถในการกำกับดูแลอัตโนมัติ (เช่น สายข้อมูลหรือการตรวจสอบคุณภาพ) อาจยังคงแข็งแกร่งที่สุดสำหรับบริการ Azure ดั้งเดิม โดยมักจะต้องมีการพัฒนา API แบบกำหนดเองสำหรับความสมบูรณ์ในการผสานรวมการไหลของข้อมูลที่ไม่ใช่ Azure [36] สิ่งนี้บ่งชี้ว่า Purview สามารถนำเสนอเส้นทางที่รวดเร็วกว่าในการสร้าง

มุมมองเมทาดาตาแบบรวมศูนย์ ทั่วทั้งสภาพแวดล้อมข้อมูลที่มีความหลากหลายอย่างแท้จริง อย่างไรก็ตาม วิศวกรข้อมูลจะต้องประเมินอย่างรอบคอบว่า "การสแกน" นี้ให้ความลึกของการกำกับดูแลที่เพียงพอหรือไม่สำหรับสินทรัพย์ข้อมูลที่สำคัญที่อยู่ภายนอก Azure หากต้องการการกำกับดูแลอัตโนมัติอย่างลึกซึ้งสำหรับแหล่งข้อมูลที่ไม่ใช่ Azure วิศวกรควรคาดการณ์ความพยายามในการ

พัฒนาแบบกำหนดเองที่คล้ายคลึงกัน (เช่น การใช้ Apache Atlas APIs) เช่นเดียวกับ Dataplex แม้ว่า Purview จะอ้างว่ามีการสแกนเริ่มต้นที่กว้างกว่าก็ตาม ซึ่งหมายถึงการทำความเข้าใจการแลกเปลี่ยนระหว่างการจัดทำแค็ตตาล็อกที่กว้างขวางและการกำกับดูแลอัตโนมัติอย่างลึกซึ้งทั่วทั้งแพลตฟอร์มที่หลากหลาย

การเปรียบเทียบโดยตรง: Dataplex กับ Purview

เพื่อให้วิศวกรข้อมูลเข้าใจความแตกต่างระหว่าง Dataplex และ Purview ได้อย่างชัดเจน การเปรียบเทียบคุณสมบัติและรูปแบบการกำหนดราคาโดยตรงจึงเป็นสิ่งสำคัญ

ตารางเปรียบเทียบคุณสมบัติ

คุณสมบัติ	Google Dataplex	Microsoft Purview
เมทาดาตาและแค็ตตาล็อก	แค็ตตาล็อกแบบรวมศูนย์สำหรับเมทาดาตาทางธุรกิจ เทคนิค และรันไทม์ ใช้ AI/ML เพื่อค้นหาความสัมพันธ์และข้อมูลเชิงลึก รองรับการค้นหาด้วยภาษาธรรมชาติ [6]	Unified Catalog และ Data Map รวบรวมเมทาดาตาจากระบบที่หลากหลาย รวมถึงการสแกนและการจัดประเภทอัตโนมัติ [7]
คุณภาพข้อมูลและการทำโปรไฟล์	กำหนดและวัดคุณภาพข้อมูลใน BigQuery ทำโปรไฟล์ข้อมูลเพื่อระบุลักษณะเฉพาะ จัดการกฎ "ในรูปแบบโค้ด" (กฎที่กำหนดไว้ล่วงหน้า, SQL กำหนดเอง) [6]	ความสามารถด้านคุณภาพข้อมูลในตัว ช่วยให้เจ้าของข้อมูลกำกับดูแลและปรับปรุงคุณภาพได้ มีข้อมูลเชิงลึกอัตโนมัติ การแก้ไขบันทึกข้อผิดพลาด [7]

สายข้อมูล	ติดตามการเคลื่อนที่และการแปลงข้อมูล สายข้อมูลอัตโนมัติสำหรับบริการ GCP (BigQuery, Dataflow, Dataproc) ขยายได้ถึงบุคคลที่สามผ่าน Data Lineage API/OpenLineage [6]	**ข้อจำกัด: เก็บข้อมูลเพียง 30 วัน** [20]	สายข้อมูลอัตโนมัติสำหรับบริการ Azure (ADF, Synapse, Power BI, Fabric) รองรับสายข้อมูลระดับสินทรัพย์และคอลัมน์ [35] สามารถสร้างสายข้อมูลแบบกำหนดเองผ่าน Apache Atlas API/REST API [36]	
การผสานรวมและการขยาย	**GCP Native:** ผสานรวมอย่างลึกซึ้งกับ BigQuery, Cloud Storage, Cloud SQL, Spanner, Vertex AI, Pub/Sub, Dataform, Dataproc Metastore [6]	**On-Prem/Multi-Cloud:** ขยายได้ผ่าน Custom Entries, Custom Connectors (PySpark), Metadata Import API [2]	**Azure Native:** ผสานรวมอย่างแข็งแกร่งกับ Azure Data Factory, Synapse Analytics, SQL DB, ADLS Gen2, Power BI, Fabric [7]	**On-Prem/Multi-Cloud:** Data Map สแกนแหล่งข้อมูลที่หลากหลาย (AWS S3, Google BigQuery) [30] มี Custom Connectors/REST APIs [46]
การจัดแนวสถาปัตยกรรม	"Intelligent Data Fabric" ที่สนับสนุนหลักการ Data Mesh (การกระจายอำนาจการเป็นเจ้าของข้อมูล, การกำกับดูแลแบบรวมศูนย์) [13]	"Federated Governance" และโมเดล "Data as a Product" (DaaP) โดยเน้นเจ้าของผลิตภัณฑ์ข้อมูลและโดเมนการกำกับดูแล [9]		
ความสามารถด้าน AI/ML	AI/ML สำหรับการค้นพบเมทาดาตา, Semantic Search, Data Insights (สร้างคำถามภาษาธรรมชาติ) [16]	AI-driven Insights สำหรับคุณภาพข้อมูล, Automated Classification, DSPM for AI [8]		

ตารางเปรียบเทียบราคาและผลกระทบด้านต้นทุน

ส่วนประกอบราคา	Google Dataplex	Microsoft Purview
หน่วยประมวลผลหลัก	**Data Compute Unit (DCU) ชั่วโมง:** Standard Processing (ค้นพบเมทาดาตา, ลงทะเบียน) มี 100 DCU ชั่วโมงฟรี/เดือน Premium Processing (DQ, Profiling, Lineage) ไม่มีฟรี และมีอัตราสูงกว่า [12]	**Data Governance Processing Unit (DGPU):** สำหรับคุณภาพข้อมูลและการจัดการสุขภาพข้อมูล ขึ้นอยู่กับประเภทกฎ ปริมาณข้อมูล และแหล่งที่มา [34]
การจัดเก็บเมทาดาตา	เมทาดาตาทางเทคนิคที่รวบรวมอัตโนมัติฟรี เมทาดาตาที่เสริม (Business Glossary, Custom Aspects) คิดค่าบริการ (1 MiB แรกฟรี จากนั้น $2/GiB/เดือน) [12]	Data Map มีหน่วยความจุ (CU) สำหรับพื้นที่จัดเก็บ (10 GB/CU) ซึ่งปรับขนาดได้อย่างยืดหยุ่น [29]
การใช้งาน API	1 ล้านครั้งแรกฟรี/เดือน หลังจากนั้น $10/100,000 ครั้ง สำหรับ Data Catalog API และ Data Lineage API [12]	การใช้งาน API สำหรับเมทาดาตาและนโยบายอาจมีค่าใช้จ่าย ขึ้นอยู่กับปริมาณการเรียกใช้ [37] (รายละเอียดเฉพาะเจาะจงน้อยกว่า Dataplex)
ระดับฟรี/การทดลองใช้	เครดิตฟรี $300 สำหรับ POC, 20+ ผลิตภัณฑ์ฟรีตลอดเวลา [16]	การทดลองใช้โซลูชันการกำกับดูแลข้อมูลฟรี (จำกัด 1,000 สินทรัพย์ที่ติดคำอธิบายประกอบ) [29]

ความแตกต่างที่สำคัญและส่วนที่ทับซ้อนกัน

- **ความลึกของการผสานรวมคลาวด์ดั้งเดิม:** Dataplex มีการผสานรวมอย่างลึกซึ้งและอัตโนมัติภายในระบบนิเวศของ GCP ซึ่งทำให้เป็นตัวเลือกที่แข็งแกร่งสำหรับองค์กรที่ใช้ GCP เป็นหลัก Purview มีความแข็งแกร่งคล้ายกันภายใน Azure และ Microsoft 365
- **แนวทางการกำกับดูแลและการเป็นเจ้าของข้อมูล:** Dataplex สอดคล้องกับแนวคิด Data Mesh โดยจัดเตรียมโครงสร้างที่รองรับการกระจายอำนาจการเป็นเจ้าของข้อมูลโดยมีศูนย์กลางการกำกับดูแล Purview เน้นโมเดล Data as a Product และแนวทางการกำกับดูแลแบบรวมศูนย์ ซึ่งส่งเสริมการจัดการข้อมูลแบบกระจายอำนาจพร้อมเครื่องมือสำหรับเจ้าของธุรกิจ
- **ความสามารถในการขยายสำหรับสถานการณ์ไฮบริด/มัลติคลาวด์:** ทั้งสองแพลตฟอร์มสามารถผสานรวมกับแหล่งข้อมูลภายนอกได้ แต่ต้องใช้ความพยายามในการวิศวกรรมแบบกำหนดเอง

อย่างมีนัยสำคัญสำหรับแหล่งข้อมูลที่ไม่ใช่คลาวด์ดั้งเดิม Dataplex ต้องการการสร้างตัวเชื่อมต่อและไปป์ไลน์การนำเข้าที่กำหนดเอง ในขณะที่ Purview ใช้ Data Map สำหรับการสแกนที่กว้างกว่า และ Apache Atlas API สำหรับสายข้อมูลแบบกำหนดเอง
- **ความสมบูรณ์ของชุดคุณสมบัติ:** Purview นำเสนอชุดโซลูชันที่ครอบคลุมมากกว่า โดยรวมความปลอดภัยของข้อมูล (DLP, Insider Risk) และการปฏิบัติตามข้อกำหนดเข้ากับการกำกับดูแล หลัก Dataplex มุ่งเน้นไปที่ Data Fabric และการกำกับดูแลข้อมูลเป็นหลัก โดยมีคุณสมบัติ AI/ML ที่แข็งแกร่งสำหรับการค้นพบและข้อมูลเชิงลึก

การเลือกแพลตฟอร์มที่เหมาะสม: คำแนะนำสำหรับวิศวกรข้อมูล

การตัดสินใจเลือกระหว่าง Google Dataplex และ Microsoft Purview ขึ้นอยู่กับบริบทและลำดับความสำคัญเฉพาะขององค์กรเป็นอย่างมาก วิศวกรข้อมูลควรพิจารณาปัจจัยต่อไปนี้อย่างรอบคอบ:

คำแนะนำตามสถานการณ์

- **สภาพแวดล้อมที่ใช้ GCP เป็นหลัก:** สำหรับองค์กรที่ลงทุนอย่างมากใน Google Cloud และมีการดำเนินงานส่วนใหญ่ใน GCP, **Google Dataplex** เป็นตัวเลือกที่เหมาะสมที่สุด การผสานรวมอย่างลึกซึ้งกับบริการ GCP ช่วยลดความซับซ้อนในการจัดการเมทาดาตา การทำโปรไฟล์ และสายข้อมูลภายในระบบนิเวศนั้นๆ นอกจากนี้ การสนับสนุนสถาปัตยกรรม Data Mesh ยังช่วยให้องค์กรสามารถนำแนวทางปฏิบัติที่ทันสมัยมาใช้ในการจัดการข้อมูลแบบกระจายอำนาจได้ง่ายขึ้น
- **สภาพแวดล้อมที่ใช้ Azure เป็นหลัก:** หากองค์กรมีโครงสร้างพื้นฐานและไปป์ไลน์ข้อมูลส่วนใหญ่อยู่ใน Microsoft Azure, **Microsoft Purview** จะเป็นตัวเลือกที่เหนือกว่า ชุดโซลูชันที่ครอบคลุมของ Purview ซึ่งรวมถึงการกำกับดูแลข้อมูล ความปลอดภัย และการปฏิบัติตามข้อกำหนด จะช่วยให้วิศวกรข้อมูลสามารถจัดการความเสี่ยงและข้อกำหนดด้านกฎระเบียบได้อย่างมีประสิทธิภาพมากขึ้นภายในแพลตฟอร์มเดียว การเน้นโมเดล Data as a Product ยังส่งเสริมการสร้างสินทรัพย์ข้อมูลที่ใช้งานได้และมีคุณภาพสูง
- **กลยุทธ์ไฮบริดหรือมัลติคลาวด์:** ทั้ง Dataplex และ Purview ต่างก็มีความสามารถในการผสานรวมกับแหล่งข้อมูลภายนอกและสภาพแวดล้อมมัลติคลาวด์ อย่างไรก็ตาม วิศวกรข้อมูลควรตระหนักว่าการผสานรวมที่ลึกซึ้งและอัตโนมัติที่สุดยังคงจำกัดอยู่เฉพาะบริการคลาวด์ดั้งเดิมของแต่ละแพลตฟอร์ม การผสานรวมแหล่งข้อมูลที่ไม่ใช่คลาวด์ดั้งเดิมมักต้องใช้ความพยายามในการวิศวกรรมแบบกำหนดเองอย่างมีนัยสำคัญ (เช่น การสร้างตัวเชื่อมต่อหรือการใช้ API) ซึ่งเพิ่มต้นทุนรวมในการเป็นเจ้าของ สำหรับองค์กรที่มีกลยุทธ์มัลติคลาวด์ที่แท้จริง การเลือกแพลตฟอร์มอาจขึ้นอยู่กับระบบคลาวด์ที่โดดเด่นที่สุด หรือความต้องการเฉพาะสำหรับการกำกับดูแลที่ครอบคลุมทั่วทั้งแพลตฟอร์มที่หลากหลาย

- **การมุ่งเน้นเฉพาะด้านคุณภาพข้อมูล สายข้อมูล หรือความปลอดภัย:**
    - หากลำดับความสำคัญสูงสุดคือ **คุณภาพข้อมูลและสายข้อมูล** ภายในสภาพแวดล้อม GCP, Dataplex มีความสามารถที่แข็งแกร่งในการทำโปรไฟล์ การกำหนดกฎคุณภาพในรูปแบบโค้ด และการติดตามสายข้อมูลอัตโนมัติ
    - หากเน้นที่ **ความปลอดภัยและการปฏิบัติตามข้อกำหนด** ที่ครอบคลุม, Purview นำเสนอชุดโซลูชันที่ผสานรวมอย่างแน่นแฟ้นสำหรับ DLP, Information Protection และ Insider Risk Management ซึ่งช่วยลดความซับซ้อนในการจัดการความเสี่ยงด้านข้อมูล

ปัจจัยที่ต้องพิจารณา

- **การลงทุนในคลาวด์ที่มีอยู่:** การเลือกแพลตฟอร์มที่สอดคล้องกับโครงสร้างพื้นฐานคลาวด์ที่มีอยู่ขององค์กรจะช่วยลดความซับซ้อนในการผสานรวมและใช้ประโยชน์จากความเชี่ยวชาญของทีมที่มีอยู่
- **ขนาดของสภาพแวดล้อมข้อมูล:** สำหรับสภาพแวดล้อมข้อมูลขนาดใหญ่มาก วิศวกรข้อมูลควรตรวจสอบข้อจำกัดของระบบ โควตา และผลกระทบต่อประสิทธิภาพการทำงานของแต่ละแพลตฟอร์มอย่างละเอียด
- **ข้อกำหนดการปฏิบัติตามข้อกำหนด:** องค์กรที่มีข้อกำหนดด้านกฎระเบียบที่เข้มงวด (เช่น HIPAA, GDPR, SOX) ควรประเมินความสามารถในการรักษาข้อมูลสายข้อมูลระยะยาว (สำหรับ Dataplex) และข้อจำกัดของนโยบาย (สำหรับ Purview) อย่างรอบคอบ
- **ความเชี่ยวชาญของทีม:** ความคุ้นเคยของทีมวิศวกรข้อมูลกับบริการคลาวด์และเครื่องมือเฉพาะของผู้จำหน่ายแต่ละรายจะส่งผลต่อเส้นทางการเรียนรู้และความเร็วในการนำไปใช้
- **งบประมาณ:** ทำความเข้าใจรูปแบบการกำหนดราคาโดยละเอียดของแต่ละแพลตฟอร์ม และพิจารณาผลกระทบของกลยุทธ์การใช้งานต่อต้นทุนรวม
- **แผนงานในอนาคต:** พิจารณาแผนงานและวิสัยทัศน์ของแต่ละแพลตฟอร์ม เพื่อให้แน่ใจว่าสอดคล้องกับกลยุทธ์ข้อมูลระยะยาวขององค์กร

บทสรุป

ทั้ง Google Dataplex และ Microsoft Purview เป็นแพลตฟอร์มการกำกับดูแลข้อมูลที่มีประสิทธิภาพสูง ซึ่งได้รับการออกแบบมาเพื่อจัดการกับความท้าทายที่ซับซ้อนของสภาพแวดล้อมข้อมูลสมัยใหม่ ทั้งสองแพลตฟอร์มนำเสนอความสามารถหลักที่สำคัญสำหรับวิศวกรข้อมูล เช่น การจัดการเมทาดาตา คุณภาพข้อมูล และสายข้อมูล อย่างไรก็ตาม จุดแข็งและแนวทางของแต่ละแพลตฟอร์มแตกต่างกันอย่างมีนัยสำคัญ

Dataplex โดดเด่นในการผสานรวมอย่างลึกซึ้งกับระบบนิเวศของ Google Cloud และการจัดแนวสถาปัตยกรรมกับหลักการ Data Mesh ซึ่งทำให้เป็นตัวเลือกที่น่าสนใจสำหรับองค์กรที่ใช้ GCP เป็นหลักและมุ่งมั่นที่จะใช้โครงสร้างข้อมูลแบบกระจายอำนาจ

ในทางกลับกัน Purview นำเสนอชุดโซลูชันที่ครอบคลุมมากขึ้น ซึ่งรวมการกำกับดูแลข้อมูล ความปลอดภัย และการปฏิบัติตามข้อกำหนดเข้าไว้ด้วยกันอย่างแน่นแฟ้น โดยเน้นที่โมเดล Data as a Product และการกำกับดูแลแบบรวมศูนย์ ทำให้เป็นตัวเลือกที่แข็งแกร่งสำหรับองค์กรที่ใช้ Azure เป็นหลักและมีข้อกำหนดด้านความปลอดภัยและการปฏิบัติตามข้อกำหนดที่เข้มงวด

สำหรับวิศวกรข้อมูล การเลือกแพลตฟอร์มที่เหมาะสมนั้นไม่ได้เกี่ยวกับว่าแพลตฟอร์มใด "ดีกว่า" แต่เกี่ยวกับว่าแพลตฟอร์มใดที่สอดคล้องกับกลยุทธ์คลาวด์ที่มีอยู่ สถาปัตยกรรมข้อมูลที่ต้องการ และลำดับความสำคัญทางธุรกิจขององค์กรมากที่สุด การทำความเข้าใจความแตกต่างทางเทคนิค รูปแบบการกำหนดราคา และความพยายามในการผสานรวมสำหรับแหล่งข้อมูลที่ไม่ใช่คลาวด์ดั้งเดิมเป็นสิ่งสำคัญในการตัดสินใจอย่างมีข้อมูล เพื่อให้มั่นใจว่าแพลตฟอร์มที่เลือกจะสามารถสนับสนุนความต้องการด้านวิศวกรรมข้อมูลและการกำกับดูแลข้อมูลขององค์กรได้อย่างมีประสิทธิภาพในระยะยาว

Works cited

1. Learn about Microsoft Purview, accessed August 4, 2025, https://docs.azure.cn/en-us/purview/purview
2. Data Quality and Governance in Google Cloud: Data Catalog vs Dataplex - dataroots, accessed August 4, 2025, https://dataroots.io/blog/data-quality-and-governance-in-google-cloud
3. Implementing Data Governance with Microsoft Purview - Emergent Software, accessed August 4, 2025, https://www.emergentsoftware.net/blog/implementing-data-governance-with-microsoft-purview/
4. Simplifying Data Management with Google Cloud Dataplex - CloudThat, accessed August 4, 2025, https://www.cloudthat.com/resources/blog/simplifying-data-management-with-google-cloud-dataplex
5. What is Microsoft Purview? Benefits, Features (2025) - Infrassist, accessed August 4, 2025, https://www.infrassist.com/microsoft-purview/
6. Dataplex Universal Catalog overview | Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/introduction
7. Learn about data governance with Microsoft Purview | Microsoft Learn, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/data-governance-overview
8. Ensuring Data Quality with Azure Purview: Features and Best Practices - XenonStack, accessed August 4, 2025, https://www.xenonstack.com/blog/data-quality-with-azure-purview
9. Learn about Microsoft Purview Unified Catalog, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/unified-catalog

10. Making AI Work: How Google Cloud Dataplex Brings Order to Your Data - Medium, accessed August 4, 2025, https://medium.com/curione-data-engineering/making-ai-work-how-google-cloud-dataplex-brings-order-to-your-data-56edcd847f16
11. Microsoft Purview and the Shift to Data as a Product, accessed August 4, 2025, https://erstudio.com/blog/microsoft-purview-and-the-shift-to-data-as-a-product/
12. Deciphering Dataplex Consumption in Google Cloud Billing | by Justin Taras - Medium, accessed August 4, 2025, https://medium.com/@jtaras/deciphering-dataplex-consumption-in-google-cloud-billing-3a260b6c8113
13. Cloud Dataplex and Data Mesh Architecture in GCP, accessed August 4, 2025, https://www.gcpstudyhub.com/pages/blog/cloud-dataplex-and-data-mesh-architecture-in-gcp
14. Google Dataplex- A Game Changer in Data Fabric Era - HCLTech, accessed August 4, 2025, https://www.hcltech.com/blogs/google-dataplex-a-game-changer-in-data-fabric-era
15. Using BigQuery Dataplex to build a data mesh | Google Cloud Blog, accessed August 4, 2025, https://cloud.google.com/blog/products/data-analytics/using-bigquery-dataplex-to-build-a-data-mesh/
16. Dataplex Universal Catalog documentation | Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs
17. Dataplex Universal Catalog | Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex
18. About data catalog management in Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/catalog-overview
19. Auto data quality overview | Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/auto-data-quality-overview
20. About data lineage | Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/about-data-lineage
21. Data Lineage API – Marketplace - Google Cloud console, accessed August 4, 2025, https://console.cloud.google.com/marketplace/product/google/datalineage.googleapis.com
22. Build a data mesh | Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/build-a-data-mesh
23. Integrate data sources with Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/integrate-data-sources
24. Manage entries and ingest custom sources | Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/ingest-custom-sources

25. Dataplex Lineage Costs - Data Analytics - Google Developer forums, accessed August 4, 2025, https://discuss.google.dev/t/dataplex-lineage-costs/193842
26. Google Dataplex Reviews 2025: Details, Pricing, & Features | G2, accessed August 4, 2025, https://www.g2.com/products/google-dataplex/reviews
27. Quotas and limits | Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/quotas
28. Best practices for Dataplex Universal Catalog - Google Cloud, accessed August 4, 2025, https://cloud.google.com/dataplex/docs/best-practices
29. Microsoft Purview | Microsoft Learn, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/
30. Best practices for scanning data sources in Microsoft Purview Data Map, accessed August 4, 2025, https://docs.azure.cn/en-us/purview/data-map-scanning-best-practices
31. Microsoft Purview data security and governance overview, accessed August 4, 2025, https://learn.microsoft.com/en-us/graph/security-datasecurityandgovernance-overview
32. Microsoft Purview: Guide to Data Governance, Compliance, and Security, accessed August 4, 2025, https://dynatechconsultancy.com/blog/microsoft-purview-data-governance-compliance-and-security
33. What's new in Microsoft Purview, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/whats-new
34. Billing in Microsoft Purview Data Governance, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/data-governance-billing
35. Data lineage - Cloud Adoption Framework - Microsoft Learn, accessed August 4, 2025, https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/govern-lineage
36. Data lineage user guide for classic Microsoft Purview Data Catalog, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/data-gov-classic-lineage-user-guide
37. Learn about Microsoft Purview collections metadata policy and roles APIs, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/legacy/tutorial-metadata-policy-collections-apis
38. Metadata driven Etl data lineage on purview - Microsoft Q&A, accessed August 4, 2025, https://learn.microsoft.com/en-us/answers/questions/690872/metadata-driven-etl-data-lineage-on-purview
39. Microsoft Purview Unified Data Governance - element61, accessed August 4, 2025, https://www.element61.be/en/competence/microsoft-purview-unified-data-governance
40. Sensitive information type limits | Microsoft Learn, accessed August 4, 2025,

https://learn.microsoft.com/en-us/purview/sit-limits
41. Limits for Microsoft 365 retention policies and retention label policies - Microsoft Learn, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/retention-limits
42. Get started with data governance experience in Microsoft Purview, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/data-governance-get-started
43. Plan for Microsoft Purview Unified Catalog with best practices, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/unified-catalog-plan
44. Microsoft Purview network architecture and best practices, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/legacy/concept-best-practices-network
45. Learn about Microsoft Purview Data Map | Microsoft Learn, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/data-map
46. Create a custom connector from scratch - Microsoft Learn, accessed August 4, 2025, https://learn.microsoft.com/en-us/connectors/custom-connectors/define-blank
47. Set up a connector to import third-party insider risk detections (preview) - Microsoft Learn, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/import-insider-risk-indicators
48. Microsoft Purview - Zetaris, accessed August 4, 2025, https://kbase.zetaris.com/knowledge/how-to-access-microsoft-purview-data-through-lightning
49. Connect to your Microsoft Purview and scan data sources privately and securely, accessed August 4, 2025, https://learn.microsoft.com/en-us/purview/data-gov-classic-private-link-end-to-end
50. Microsoft Purview Data Lifecycle Management Pros and Cons | User Likes & Dislikes - G2, accessed August 4, 2025, https://www.g2.com/products/microsoft-purview-data-lifecycle-management/reviews?qs=pros-and-cons