central data TECHNOLOGY
a member of CTI Group

Memahami AWS Glue

Topik

Apa Itu AWS Glue?

AWS Glue adalah layanan *serverless data integration* dari Amazon Web Services yang dirancang untuk membantu organisasi dalam mengelola proses integrasi data secara *end-to-end* tanpa perlu mengelola infrastruktur. Melalui pendekatan *serverless*, AWS Glue memungkinkan pengguna untuk secara otomatis menemukan, mempersiapkan, dan menggabungkan data yang berasal dari berbagai sumber, baik dari lingkungan *on-premises* maupun layanan *cloud*.

AWS Glue berperan penting dalam membangun *data pipeline* modern yang digunakan untuk kebutuhan analitik, *business intelligence*, serta aplikasi *machine learning (ML)*. Layanan ini mendukung proses *extract, transform, and load (ETL)* maupun *extract, load, and transform (ELT)* sehingga data dapat diolah dan disiapkan sesuai dengan kebutuhan analisis.

Salah satu komponen inti AWS Glue adalah *AWS Glue Data Catalog*, yang berfungsi sebagai repositori metadata terpusat. *Data Catalog* menyimpan informasi mengenai skema, tabel, dan lokasi data, sehingga dapat digunakan bersama oleh berbagai layanan AWS lainnya. Selain itu, AWS Glue memungkinkan pengguna untuk membangun dan menjalankan *ETL workflow* secara otomatis, terjadwal, atau berbasis *event*, tanpa perlu melakukan *provisioning*, konfigurasi, maupun pemeliharaan *server* secara manual.

Keunggulan Utama AWS Glue

AWS Glue menawarkan sejumlah keunggulan yang menjadikannya solusi integrasi data yang efisien dan *scalable* bagi organisasi modern.

Pendekatan Serverless

Sebagai layanan *serverless*, AWS Glue menghilangkan kebutuhan untuk melakukan *provisioning* dan pengelolaan sumber daya komputasi secara manual. AWS secara otomatis menangani proses penskalaan sumber daya sesuai dengan beban kerja yang dijalankan, sehingga tim dapat lebih fokus pada pengolahan data dan logika bisnis dibandingkan operasional infrastruktur.

Dukungan Multi-Sumber Data

AWS Glue mendukung integrasi data dari berbagai sumber melalui beragam *connector*, termasuk data yang tersimpan di Amazon S3, *database* relasional, *data warehouse*, serta sumber data pihak ketiga. Fleksibilitas ini memungkinkan organisasi untuk mengonsolidasikan data dari berbagai sistem ke dalam satu arsitektur *data lake* atau platform analitik terpadu.

Opsi ETL Visual dan Berbasis Kode

Untuk memenuhi kebutuhan pengguna dengan latar belakang yang berbeda, AWS Glue menyediakan dua pendekatan utama dalam membangun pipeline ETL. Pengguna dapat memanfaatkan *AWS Glue Studio* dengan antarmuka visual berbasis *drag-and-drop*, atau menggunakan skrip berbasis *Python* maupun *Scala* untuk kontrol dan fleksibilitas yang lebih tinggi. Pendekatan ini mempermudah pengembangan pipeline baik untuk pengguna teknis maupun non-teknis.

Integrasi Mendalam dengan Layanan AWS

AWS Glue terintegrasi secara native dengan berbagai layanan AWS lainnya seperti *Amazon Athena*, *Amazon Redshift Spectrum*, *Amazon EMR*, dan layanan analitik lainnya. Integrasi ini memungkinkan data yang telah diproses oleh AWS Glue untuk langsung digunakan dalam analisis, pelaporan, maupun pemodelan *machine learning*, tanpa perlu proses integrasi tambahan yang kompleks.

Komponen Utama AWS Glue

AWS Glue Data Catalog

AWS Glue Data Catalog berfungsi sebagai repositori metadata terpusat yang menyimpan informasi penting terkait struktur dan lokasi data. Data Catalog menyimpan definisi tabel, skema data, koneksi ke sumber data, serta versi skema yang memungkinkan pengelolaan perubahan struktur data dari waktu ke waktu. Dengan pendekatan terpusat ini, Data Catalog menjadi fondasi utama dalam arsitektur data integration dan data lake di AWS.

Metadata yang tersimpan di dalam Data Catalog dapat digunakan secara langsung oleh berbagai layanan analitik AWS seperti Amazon Athena dan Amazon EMR. Hal ini memungkinkan pengguna untuk melakukan *query* dan analisis data tanpa perlu mendefinisikan ulang struktur data

di setiap layanan. Dengan demikian, konsistensi skema dan efisiensi akses data dapat terjaga di seluruh ekosistem AWS.

Beberapa fitur utama yang menjadikan AWS Glue Data Catalog sebagai komponen penting dalam pengelolaan data antara lain:

1) *Schema management* dan *schema evolution*, yang memungkinkan perubahan struktur data dilakukan secara terkontrol tanpa mengganggu *pipeline* yang sudah berjalan.
2) Metadata terpusat yang dapat digunakan lintas layanan AWS, sehingga memudahkan integrasi dan interoperabilitas antar layanan analitik dan pemrosesan data.
3) Statistik kolom yang digunakan untuk membantu optimasi performa kueri, terutama saat data diakses melalui layanan analitik seperti Amazon Athena.

Crawlers

AWS Glue Crawlers merupakan mekanisme otomatis untuk menemukan dan mengklasifikasikan data yang tersimpan di berbagai sumber. *Crawler* bekerja dengan cara melakukan *scan* sumber data, membaca struktur file atau tabel, lalu mengidentifikasi skema data secara otomatis. Hasil inferensi tersebut kemudian disimpan ke dalam AWS Glue Data Catalog sebagai *metadata* yang siap digunakan.

Crawler dapat dijalankan dalam berbagai mode sesuai kebutuhan operasional. Pengguna dapat menjalankannya secara *on-demand* untuk kebutuhan *ad-hoc*, menjadwalkannya secara berkala untuk menjaga konsistensi *metadata*, atau mengaktifkannya berdasarkan peristiwa tertentu melalui *event-based triggers*. Fleksibilitas ini memungkinkan perusahaan memastikan bahwa *metadata* selalu selaras dengan kondisi data aktual, terutama dalam lingkungan dengan perubahan data yang cepat dan dinamis.

Jobs & Triggers

Dalam AWS Glue, *Jobs* merepresentasikan proses eksekusi logika *extract*, *transform*, *and load* (ETL) yang bertugas mengekstrak data dari sumber, melakukan transformasi sesuai kebutuhan bisnis, dan memuat data ke tujuan akhir. *Jobs* dapat dibuat menggunakan pendekatan visual melalui AWS Glue Studio atau dengan menulis skrip berbasis Python maupun Scala untuk skenario yang lebih kompleks dan kustom.

Sementara itu, *Triggers* digunakan untuk mengotomatisasi eksekusi *Jobs* berdasarkan kondisi tertentu. *Trigger* dapat dikonfigurasi untuk menjalankan *Jobs* berdasarkan jadwal waktu, dependensi antar *Jobs*, maupun *event* tertentu. Dengan kombinasi *Jobs* dan *Triggers*, AWS Glue memungkinkan pembangunan ETL workflow yang terstruktur, terotomatisasi, dan mudah dikelola tanpa memerlukan orkestrasi eksternal.

Arsitektur *Serverless* AWS Glue

AWS Glue dibangun di atas arsitektur *serverless* yang memungkinkan seluruh pemrosesan data berjalan tanpa membutuhkan pengelolaan infrastruktur secara manual. Layanan ini menggunakan *distributed processing engine* berbasis *Apache Spark* atau *Ray* untuk mengeksekusi pekerjaan pemrosesan data dalam skala besar. Melalui pendekatan ini, pengguna tidak perlu melakukan *provisioning*, konfigurasi, maupun pemeliharaan *server* atau klaster komputasi.

Dalam arsitektur *serverless* AWS Glue, sumber daya komputasi akan dialokasikan secara dinamis berdasarkan karakteristik dan kompleksitas pekerjaan yang dijalankan. Sistem secara otomatis melakukan *auto scaling* untuk menyesuaikan kapasitas komputasi, baik saat beban kerja meningkat maupun menurun. Hal ini memastikan performa pemrosesan data tetap optimal sekaligus menjaga efisiensi biaya karena pengguna hanya membayar sumber daya yang benar-benar digunakan.

Selain itu, AWS Glue mengelola seluruh aspek operasional seperti *job orchestration*, *resource management*, dan *fault tolerance*. Jika terjadi kegagalan pada sebagian proses, sistem akan menangani pemulihan secara otomatis tanpa diperlukannya intervensi manual. Pendekatan ini menjadikan AWS Glue sangat sesuai untuk kebutuhan *data integration*, *batch processing*, maupun *streaming ETL* dalam lingkungan data modern yang membutuhkan skalabilitas tinggi, ketersediaan, dan keandalan operasional.

Fitur & Kemampuan AWS Glue

AWS Glue menyediakan rangkaian fitur komprehensif yang dirancang untuk menyederhanakan proses integrasi data sekaligus meningkatkan skalabilitas dan efisiensi operasional. Fitur-fitur ini membedakan AWS Glue dari solusi *extract, transform, and load (ETL)* tradisional yang umumnya memerlukan pengelolaan infrastruktur dan konfigurasi manual yang kompleks.

Integrasi Data Otomatis

AWS Glue menyediakan kemampuan integrasi data otomatis melalui penggunaan *crawlers* yang berfungsi untuk mendeteksi struktur dan skema data secara otomatis. Crawler akan melakukan pemindaian terhadap sumber data, kemudian menentukan skema data dan menuliskannya langsung ke dalam *AWS Glue Data Catalog*. Proses ini secara signifikan

mempercepat tahap *data discovery* dan mengurangi ketergantungan pada proses manual dalam pendefinisian *metadata*, terutama pada lingkungan data yang berskala besar dan terus berkembang.

Transformasi Data dan ETL *Visual*

Melalui *AWS Glue Studio*, pengguna dapat membangun *ETL pipeline* menggunakan antarmuka visual berbasis *drag-and-drop*. Pendekatan ini memungkinkan perancangan alur transformasi data tanpa harus menulis seluruh kode secara manual, sehingga mempermudah pengembangan pipeline bagi pengguna dengan berbagai tingkat keahlian teknis. Meskipun demikian, AWS Glue tetap memberikan fleksibilitas untuk menyesuaikan transformasi dengan kode berbasis *Python* atau *Scala* bila diperlukan, sehingga cocok untuk skenario transformasi data yang kompleks.

***Streaming* ETL**

Selain mendukung pemrosesan data berbasis *batch*, AWS Glue juga mendukung pemrosesan data berkelanjutan melalui *AWS Glue Streaming Jobs*. Fitur ini memungkinkan organisasi memproses data *real-time* dari sumber seperti *Amazon Kinesis* dan *Apache Kafka*. Dengan dukungan *streaming ETL*, data dapat ditransformasikan dan disiapkan secara terus-menerus untuk kebutuhan analitik, pemantauan operasional, maupun aplikasi yang membutuhkan *near real-time insights*.

Pengelolaan Skema dan Validasi Data

AWS Glue Schema Registry menyediakan mekanisme untuk mengelola, memvalidasi, dan melacak evolusi skema data, khususnya pada alur data *streaming*. Dengan adanya validasi skema, sistem dapat memastikan bahwa data yang masuk tetap kompatibel dengan skema yang telah ditentukan. Hal ini membantu mencegah kegagalan pemrosesan akibat perubahan struktur data yang tidak terkontrol, sekaligus menjaga konsistensi data seiring waktu.

***Machine Learning* untuk Pembersihan Data**

AWS Glue memanfaatkan kemampuan *machine learning* untuk meningkatkan kualitas data melalui fitur *FindMatches*. Fitur ini membantu mendeteksi dan mengelompokkan data duplikat atau data yang serupa tanpa memerlukan keahlian *machine learning* yang mendalam dari pengguna. Dengan pendekatan ini, proses *data cleansing* menjadi lebih efisien dan akurat, sehingga data yang digunakan untuk analitik dan pelaporan memiliki kualitas yang lebih baik.

Autoscaling

AWS Glue secara otomatis melakukan *autoscaling* dengan menambah atau mengurangi sumber daya komputasi sesuai dengan beban kerja yang dijalankan. Mekanisme ini memastikan bahwa pipeline ETL dapat berjalan dengan performa optimal pada berbagai skala data, sekaligus menjaga efisiensi biaya karena pengguna hanya membayar sumber daya yang digunakan. Pendekatan ini menjadikan AWS Glue sangat sesuai untuk beban kerja yang bersifat dinamis dan tidak terprediksi.

Notebook dan *Interactive Sessions*

AWS Glue menyediakan *serverless notebooks* dan *interactive sessions* yang memungkinkan pengguna melakukan eksplorasi data, pengujian transformasi, serta pengembangan *pipeline* secara interaktif. Fitur ini sangat berguna dalam tahap *data exploration* dan *development*, karena pengguna dapat langsung menulis dan menjalankan kode tanpa perlu menyiapkan lingkungan komputasi terpisah. Dengan demikian, siklus pengembangan pipeline data menjadi lebih cepat dan iteratif.

Manfaat Bisnis dan Teknis

AWS Glue tidak hanya memberikan kemudahan dari sisi teknis, tetapi juga menghadirkan manfaat bisnis yang signifikan bagi organisasi. Dengan pendekatan *serverless* dan integrasi mendalam dalam ekosistem AWS, AWS Glue membantu meningkatkan efisiensi operasional, skalabilitas sistem, serta fleksibilitas dalam pengelolaan data untuk berbagai kebutuhan analitik dan *artificial intelligence / machine learning (AI/ML)*.

Efisiensi Operasional

Dengan arsitektur *serverless*, AWS Glue menghilangkan kebutuhan bagi tim teknologi informasi untuk melakukan pengelolaan infrastruktur seperti provisioning server, konfigurasi klaster, maupun pemeliharaan sistem secara rutin. Seluruh aspek operasional, termasuk pengelolaan sumber daya komputasi dan orkestrasi pekerjaan ETL, ditangani secara otomatis oleh AWS. Hal ini memungkinkan tim TI dan data untuk lebih fokus pada aktivitas bernilai tambah, seperti pengembangan logika transformasi data, peningkatan kualitas data, serta analisis yang mendukung pengambilan keputusan bisnis.

Skalabilitas dan Fleksibilitas

AWS Glue dirancang untuk menangani berbagai skala beban kerja, mulai dari volume data yang relatif kecil hingga pemrosesan data dalam jumlah besar. Layanan ini secara otomatis melakukan *scaling* sesuai kebutuhan tanpa memerlukan perubahan atau penyesuaian arsitektur. Fleksibilitas ini memungkinkan organisasi untuk menyesuaikan kapasitas pemrosesan data seiring pertumbuhan bisnis, tanpa harus melakukan investasi awal yang besar atau migrasi sistem yang kompleks.

Integrasi yang Kuat

AWS Glue terintegrasi secara native dengan berbagai layanan analitik dan *data lake* di AWS, seperti *Amazon S3*, *Amazon Athena*, *Amazon Redshift*, serta layanan *machine learning*. Integrasi yang kuat ini memungkinkan data yang telah diproses oleh AWS Glue untuk langsung dimanfaatkan dalam berbagai *use case*, termasuk analitik deskriptif, pelaporan bisnis, hingga pengembangan model *AI/ML*. Dengan ekosistem yang terintegrasi, AWS Glue menjadi komponen kunci dalam membangun arsitektur data modern yang *end-to-end*.