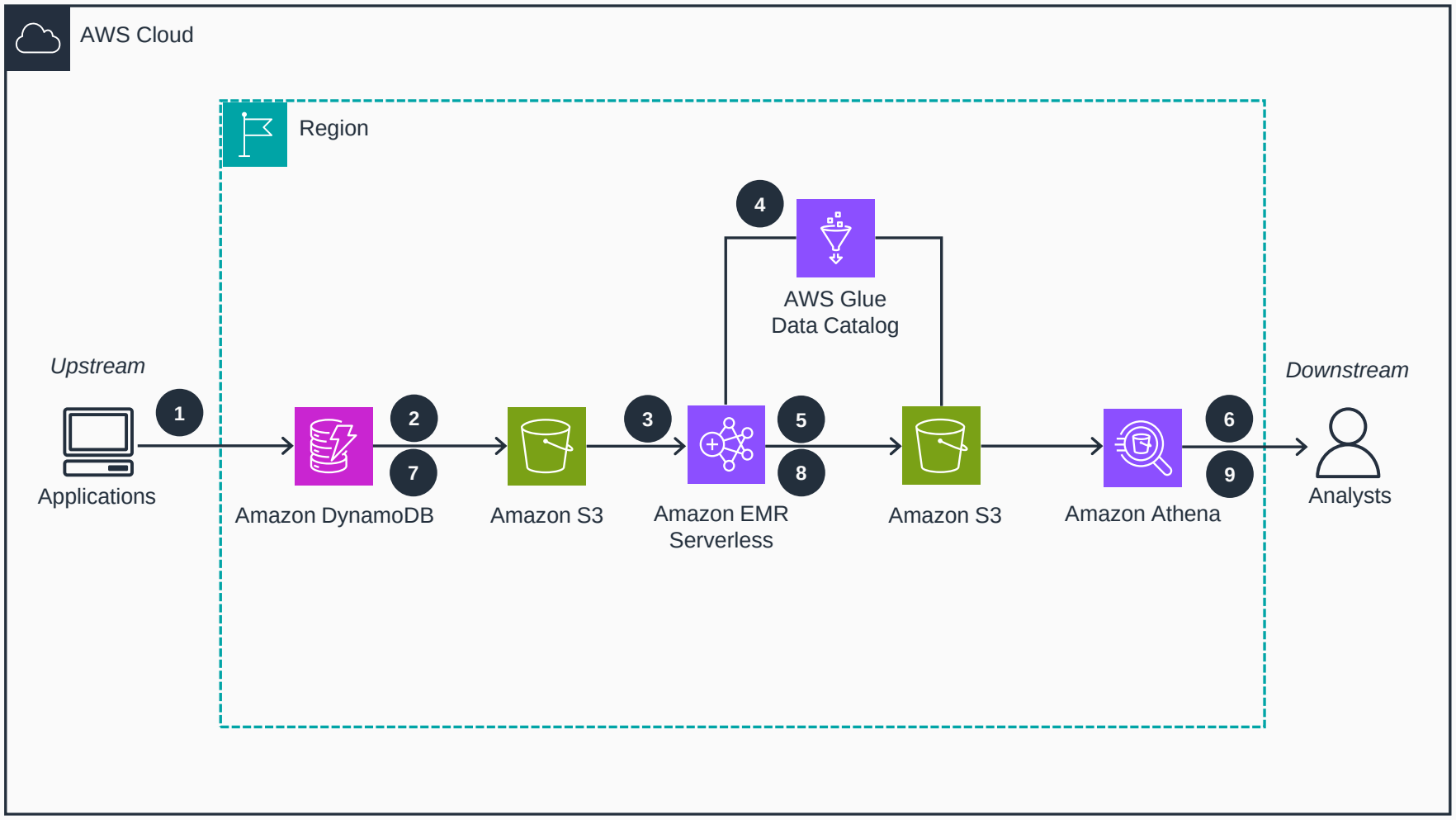
Guidance for Incremental Data Exports on AWS

This architecture diagram shows how to make incremental exports to update your downstream systems regularly using only the changed data.



1. Application traffic consistently adds, updates, and deletes items in an **Amazon DynamoDB** table.
2. Perform a full export of your **DynamoDB** table. This will write the exported data into **Amazon Simple Storage Service (Amazon S3)** in JSON format.
3. Create, prepare, and use **Amazon EMR Serverless** to read the full export of the **DynamoDB** table from **Amazon S3**. **EMR Serverless** will dynamically identify Iceberg table schema with the full set of columns that will map to all the unique attributes from your full **DynamoDB** exported dataset.
4. Create **AWS Glue Data Catalog** to persist the Iceberg table meta store and query the table from **Amazon Athena** (or any Hive Meta store compatible query engine) using the same **Glue Catalog**.
5. Use **EMR Serverless** to build the Iceberg table based on the full export of the **DynamoDB** table, and use the Iceberg table generated schema.
6. Analysts can use an **Athena** query to verify that the Iceberg table is accessible and readable. This involves running a SELECT query on the Iceberg table through **Athena** to confirm that data can be retrieved successfully and accurately.
7. Perform an incremental export of your **DynamoDB** table in JSON format. This will only export the changed data from the **DynamoDB** table since the last full or incremental export.
8. Use **EMR Serverless** to update a previously created Iceberg table with the incremental export of the **DynamoDB** table data.
9. Analysts will use the same **Athena** query to verify that the Iceberg table shows changed records. For example, if you've added or deleted items in the **DynamoDB** table after full export, the count should reflect this.

Reviewed for technical accuracy June 4, 2024

AWS Reference Architecture