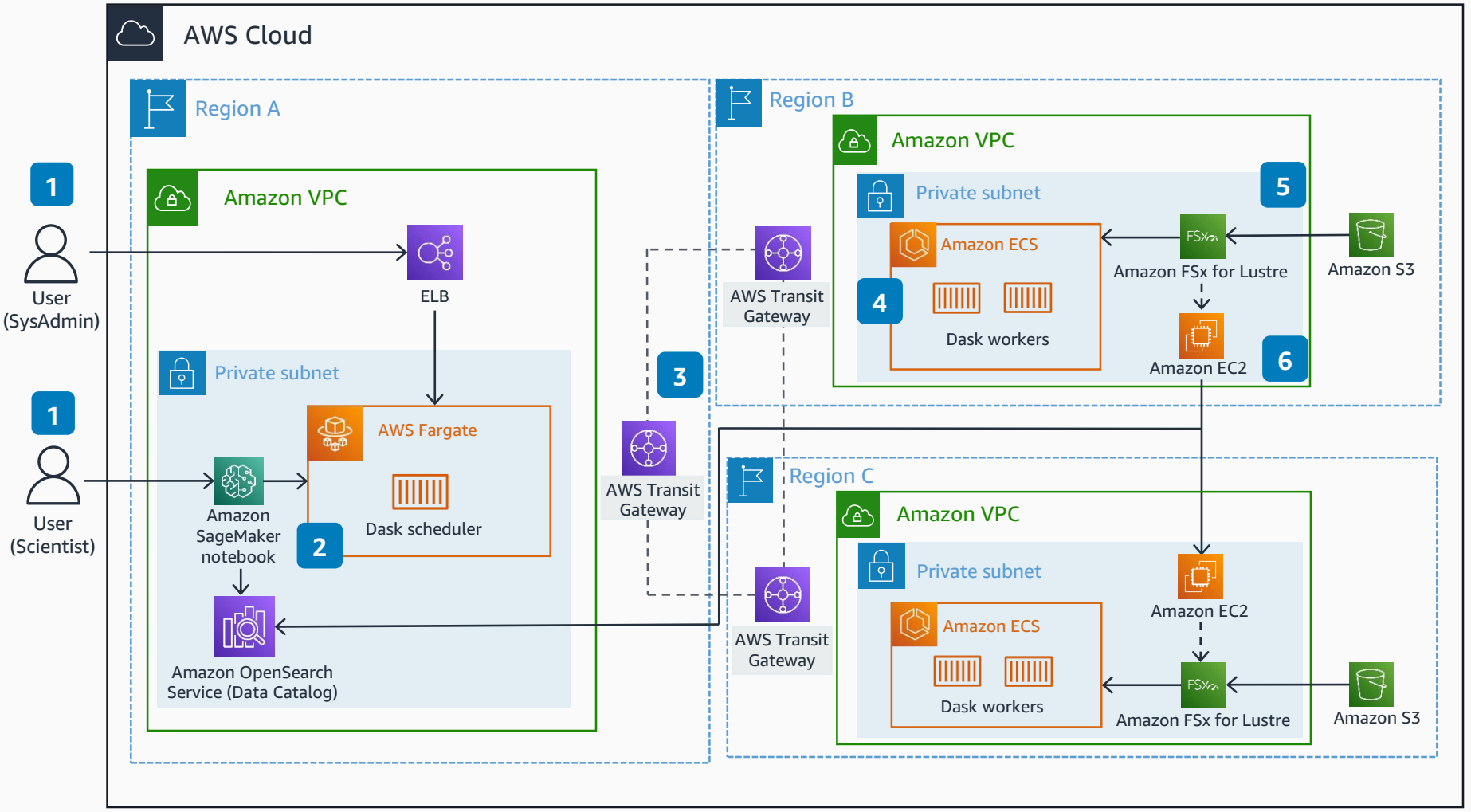
Guidance for Distributed Computing with Cross Regional Dask on AWS

This architecture shows how to perform data-proximate compute on large datasets located across multiple Regions using cross regional Dask clusters on AWS. It also minimizes cross regional traffic and your associated carbon footprint.



1. This architecture contains two personas. Through **Elastic Load Balancing (ELB)**, the system administrator (Persona 1), has access to the Dask dashboard running on the Dask scheduler. Persona 2, the scientist, accesses the **Amazon SageMaker** notebook through the AWS console.

2. **SageMaker** notebook connects to the scheduler, and the user looks up Dask workers closest to the datasets by querying the data catalog on **Amazon OpenSearch Service** and then initiating the compute request.

3. **AWS Transit Gateway** routes traffic between the Dask scheduler and Dask workers running in different AWS Regions.

4. Dask workers running as **Amazon Elastic Container Service (Amazon ECS)** tasks perform the requested compute on datasets mounted through **Amazon FSx for Lustre**. **Amazon ECS** automatically scales up Dask worker instances based on CPU usage.

5. Metadata of datasets are synced periodically from public **Amazon Simple Storage Service (Amazon S3)** buckets that are part of the **Open Data on AWS Initiative**.

6. Synced datasets are automatically indexed using an **Amazon Elastic Compute Cloud (Amazon EC2)** instance that executes a daily cronjob (a command for scheduled tasks) to update the data catalog on **OpenSearch Service**.

Reviewed for technical accuracy March 13, 2023

AWS Reference Architecture