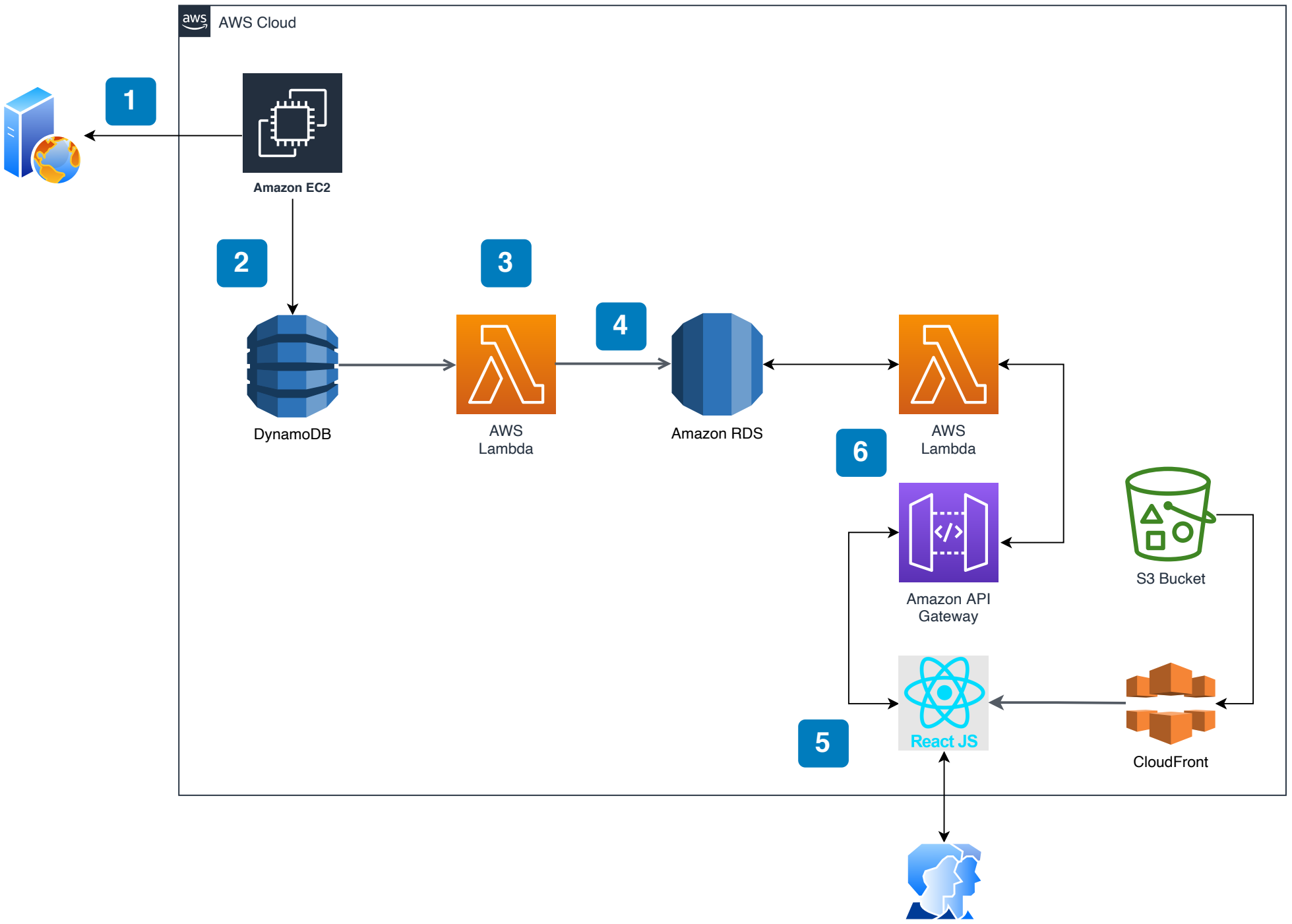
Purchase Power Parities

Architecture Diagram



1. Web parsing tool grabs a few sample records from different online market locations for PoC. Production solution would need to establish a data sharing mechanism to build out larger product sample.

2. Sample products are stored in DynamoDB using attributes that are returned from each online sore. Each product set exists under a unique key that corresponds to the product search query along with arrays of sample products under each vendor.

3. Categorization process attempts to determine if a product description can be parsed for information regarding the nature of the product. This is done using various parts of speech and using Spacy's Okapi similarity for string vectors. If a product is deemed to be similar enough to an existing subcategory in RDS, the product is matched and placed in that category. If no category is similar enough is found, the program auto generates a new category.

4. With a set of both standardized categories and NLP generated categories the system now uses quantulum3 and pint Python libraries to extract quantities from the product description. Then using NLP brand awareness and bag of words technique the string is broken down into key components to attempt to identify similar products.

5. With a set of both standardized categories and NLP generated categories and products places within those categories, a web interface provides a superset of categories with products. User can drill into categories to locate products for matching.

6. The web application allows users to confirm 2 products are identical or recategorize a product that was inaccurately labeled. Using API Gateway and Lambda functions the system records matches that can be used in PPP calculations.