Recommended Big Data Datasets

**1. Public Datasets from Kaggle**

* **Dataset**: [New York City Taxi Trip Data](https://www.kaggle.com/datasets/).
* **Description**: Large dataset with millions of rows detailing taxi trips in NYC, including timestamps, locations, and fares.
* **Size**: 20 GB+.
* **Use Case**: Demonstrates filtering, aggregations, and geospatial analysis.

**2. Amazon AWS Open Data Registry**

* **Dataset**: [Common Crawl](https://registry.opendata.aws/commoncrawl/).
* **Description**: Open dataset of web crawl data for text analysis and big data processing.
* **Size**: Petabytes (you can extract a manageable subset).
* **Use Case**: Text analytics, word count, and distributed processing.
* **Dataset**: [US Census Data](https://registry.opendata.aws/2020-census-summary-file/).
* **Description**: Large demographic dataset.
* **Size**: Varies by subset.
* **Use Case**: Visualization and demographic insights.

**3. Google BigQuery Public Datasets**

* **Dataset**: [COVID-19 Open Data](https://console.cloud.google.com/marketplace/details/bigquery-public-data/covid19-open-data).
* **Description**: Comprehensive dataset of COVID-19 cases worldwide.
* **Size**: 10 GB+.
* **Use Case**: Trend analysis, time-series forecasting.

**4. UCI Machine Learning Repository**

* **Dataset**: [Online Retail Dataset](https://archive.ics.uci.edu/ml/index.php).
* **Description**: Transactional dataset from an online retailer.
* **Size**: Medium-large (~500 MB to 1 GB).
* **Use Case**: User behavior analysis, purchase trends.

**5. Microsoft’s Open Datasets**

* **Dataset**: [US Accidents Dataset](https://azure.microsoft.com/en-us/services/open-datasets/).
* **Description**: A detailed record of traffic accidents in the U.S.
* **Size**: 1 GB+.
* **Use Case**: Geospatial and temporal analytics.

**6. OpenStreetMap Data**

* **Dataset**: [OSM Data Extracts](https://www.openstreetmap.org/).
* **Description**: Open geospatial dataset with information on maps and geographic objects.
* **Size**: Variable, depending on the region.
* **Use Case**: Spatial joins, mapping, and geospatial processing.

**7. Large-Scale Sentiment Analysis Dataset**

* **Dataset**: [IMDb Reviews Dataset](https://ai.stanford.edu/~amaas/data/sentiment/).
* **Description**: Sentiment-labeled text data for natural language processing.
* **Size**: ~5 GB.
* **Use Case**: Sentiment analysis, text classification.

**8. Open Energy Data**

* **Dataset**: [Global Power Plant Database](https://datasets.wri.org/dataset/globalpowerplantdatabase).
* **Description**: Information on power plants worldwide.
* **Size**: Medium (~500 MB to 2 GB).
* **Use Case**: Energy trend analysis, capacity forecasting.