Data Prep – Spark Workflow

Raw data
UTM Zone 10
Longitude [-126, -120]
36 files, csv, 125 GB

↓ Filter by lat, lon

Puget Sound
Lat [47.00, 49.45]
Lon [-123.75, -122.00]

→ Split →

Vessel Motion

→ Sort by time →

Time-sorted Motion

Aggregate by Device ID (mmsi) →

Time series array of **Vessel Tracks** for each vessel (3 years)

Save as parquet ↓

3 years data for modeling after data exploration

Filter to 3 months →

Exploratory data for use on a PC

Vessel Metadata, csv

Convert UTC to local time. Then partition ↓

Partitioned on year, month, day

Save as parquet ↓

Save on Blob Storage for querying