Overall:

Apache Airflow – Trigger pipeline.

Data preparation:

Marqueez or Openlineage orsimilar tool – Capture Data lineage & meta data.

Pandas - suitable for 200k

Polars - suitable for 2-10m

Dask - suitable for 10-50m

Apache Spark – Big data processing – any one

DVC – for data version control

Data warehouse (probably cloud or on premise) – for storing historical data

Model Development:

Pycaret or Azure auto ml etc – for automatic model development

Any plotting library – for model explainability and metrics

Optuna – Hyperparameter tuning

Model Deployment:

- **CI/CD:**
  - Docker – containerization
  - Kubernetes, KServe – Container orchestration
  - GitHub Actions or similar – to trigger CI/CD
- Inference Endpoint:
  - Kafka for data streaming
- MLOPS:
  - Weights and bias or Kubeflow or mlflow for experiment tracking anf model registry.
  - Evidently AI or Prometheus & Grafana to monitor performance & drift
- Inference API:
  - Fast API or Triton from nvidia