OT-6

« Large-scale cloud services for big data management »

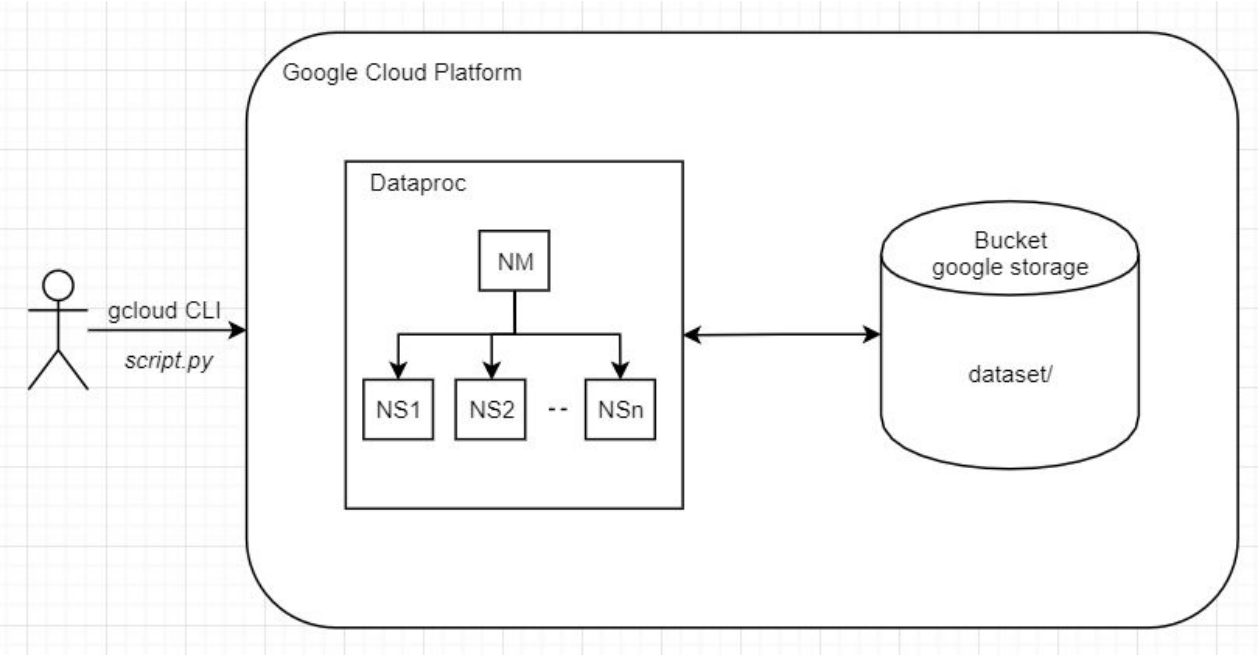
GUITTAT Clément - VIRY Baptiste

Dataset utilisé

YearPredictionMSD Data Set :

- 515345 lignes
- 90 attributs
- 1 valeur à prédire : l’année

Architecture



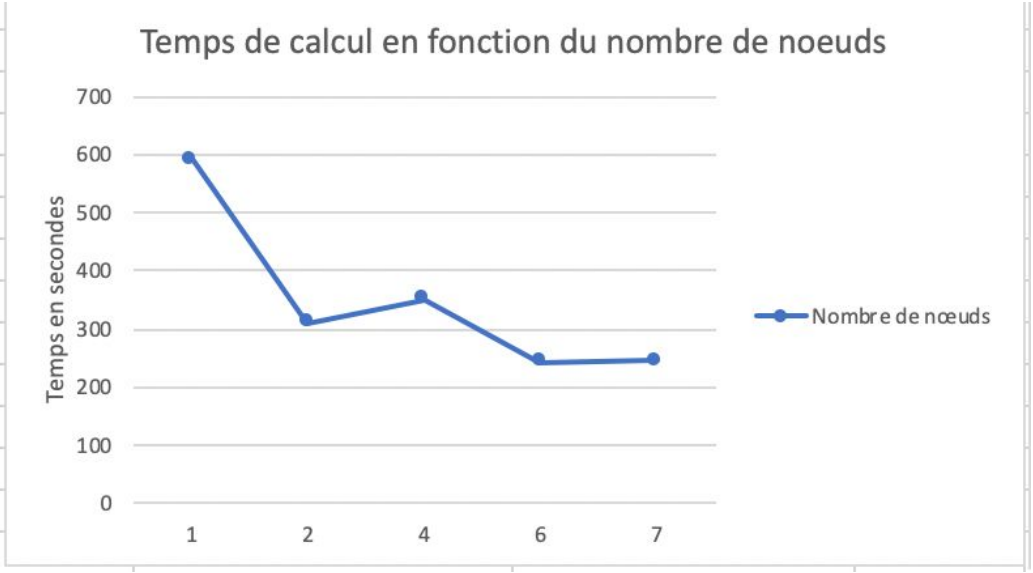
Google Cloud Platform

- Accès CLI ou interface graphique

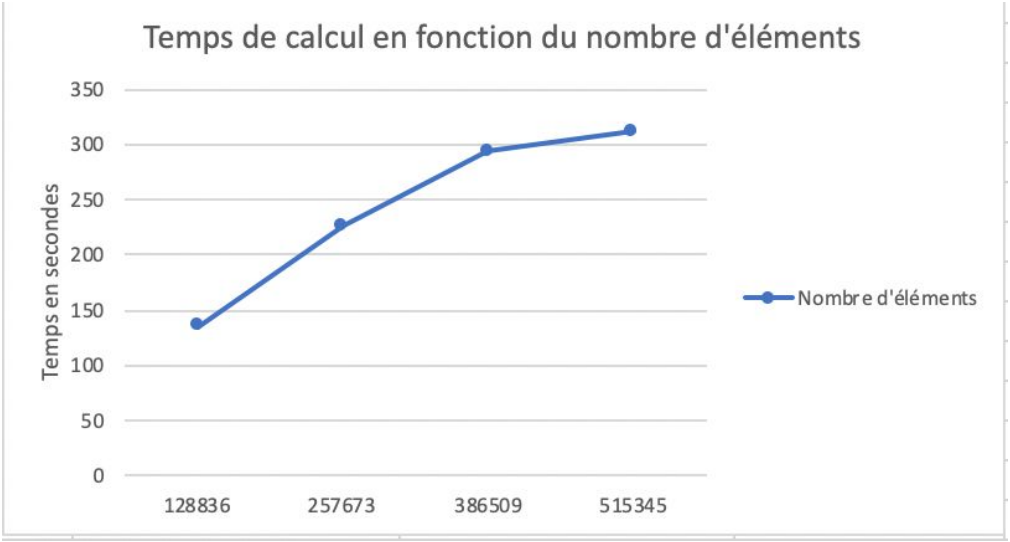
gcloud dataproc jobs submit pyspark job_full_dataset.py --cluster=cluster-1

- Stockage du dataset dans un bucket

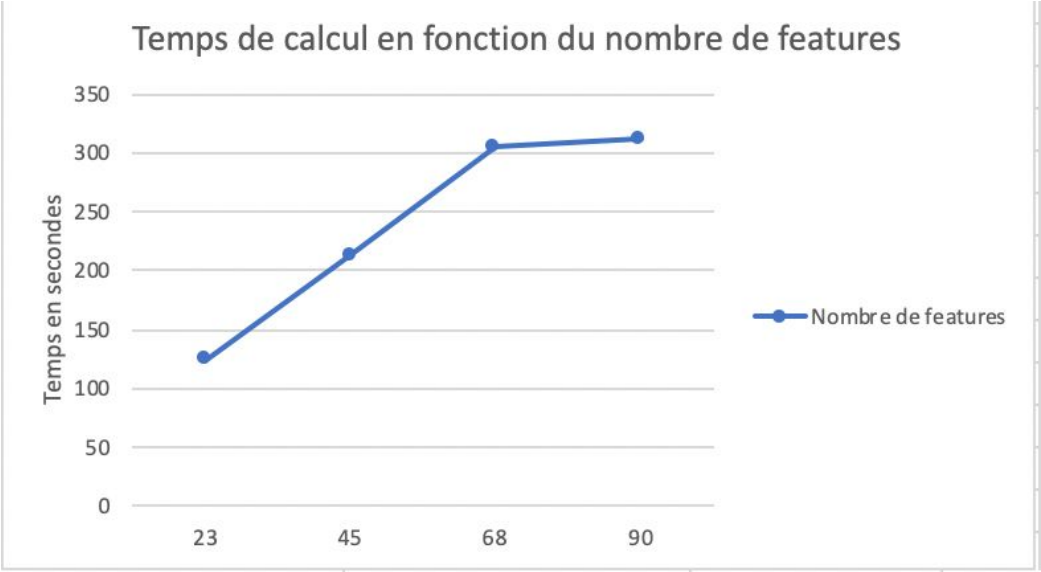
Performances selon le nombre de noeuds de calcul



Performances selon le nombre de lignes du dataset



Performances selon le nombre de colonnes du dataset



Valeur de l'erreur de notre algorithme selon le nombre de lignes du dataset