Apache Arrow Meetup 2019

TensorFlow + BigQuery Storage API + Apache Arrow

SENSY株式会社 漆山和樹 @KUrushi_ml

人生
初登壇

自己紹介

漆山和樹

SENSY株式会社
- Researcher (Data Scientist) 研究とかPoCとか
- 機械学習エンジニア

弊社のやっていること

リテイルに機械学習モデルの導入(PoC)

感性学習をするAIの研究・開発

PoCでのよく使われるもの

Storage
Analytics

BigQuery

Cloud
Storage

dmlc XGBoost

TensorFlow

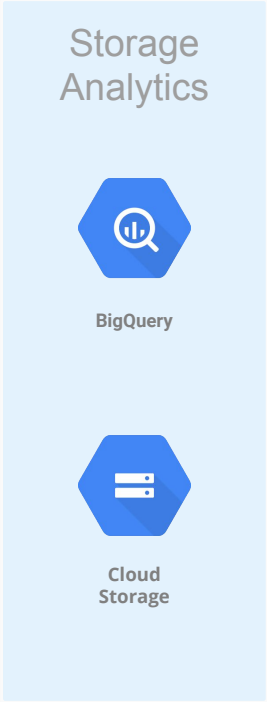
SENSY

jupyter

Compute
Engine

PoCでの課題

1. 数百GB ~ 数TBのデータ量
2. Pandasの利用
3. 密ベクトル

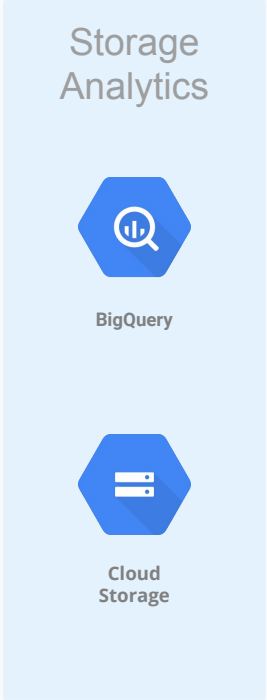


PoCでの課題

1. 数百GB ~ 数TBのデータ量
2. Pandasの利用
3. 密ベクトル

PoCでの課題

1. 数百GB ~ 数TB のデータ量
2. Pandasの利用
3. 密ベクトル

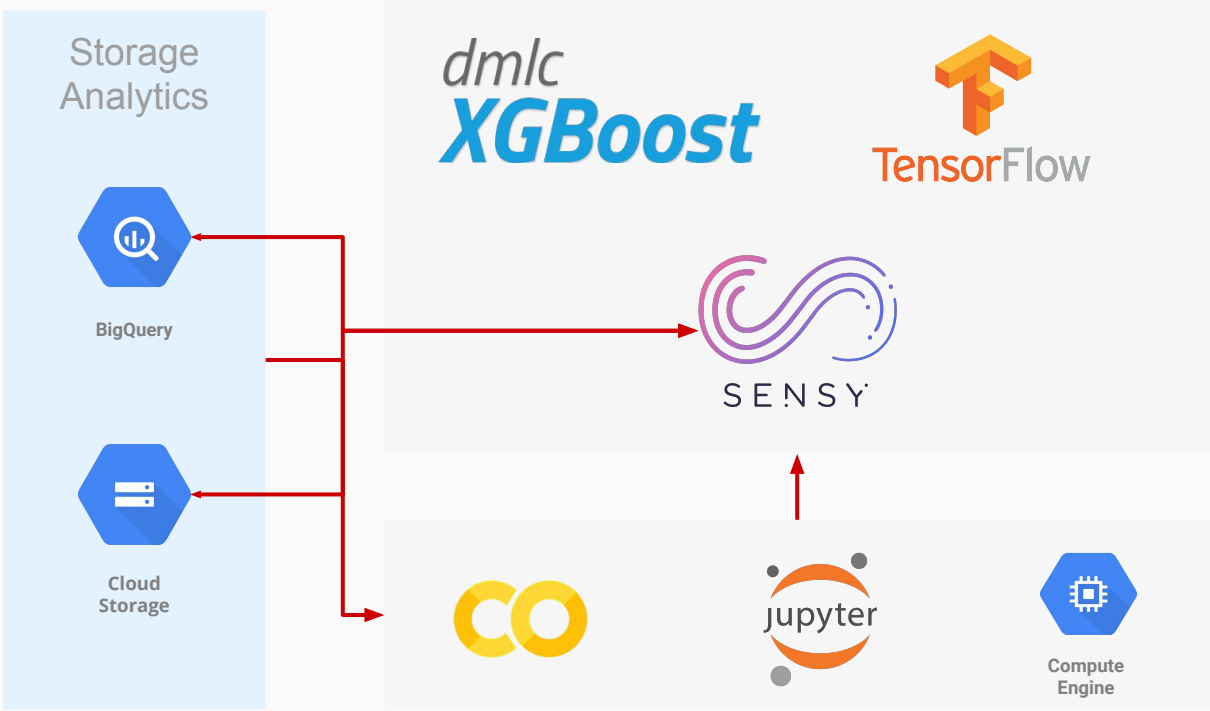


PoCでの課題

1. 数百GB ~ 数TBのデータ量
2. Pandasの利用
3. 密ベクトル

PoCでの課題

1. 数百GB ~ 数TBのデータ量
2. Pandasの利用
3. 密ベクトル



選ばれたのは、
Apache Arrowでした

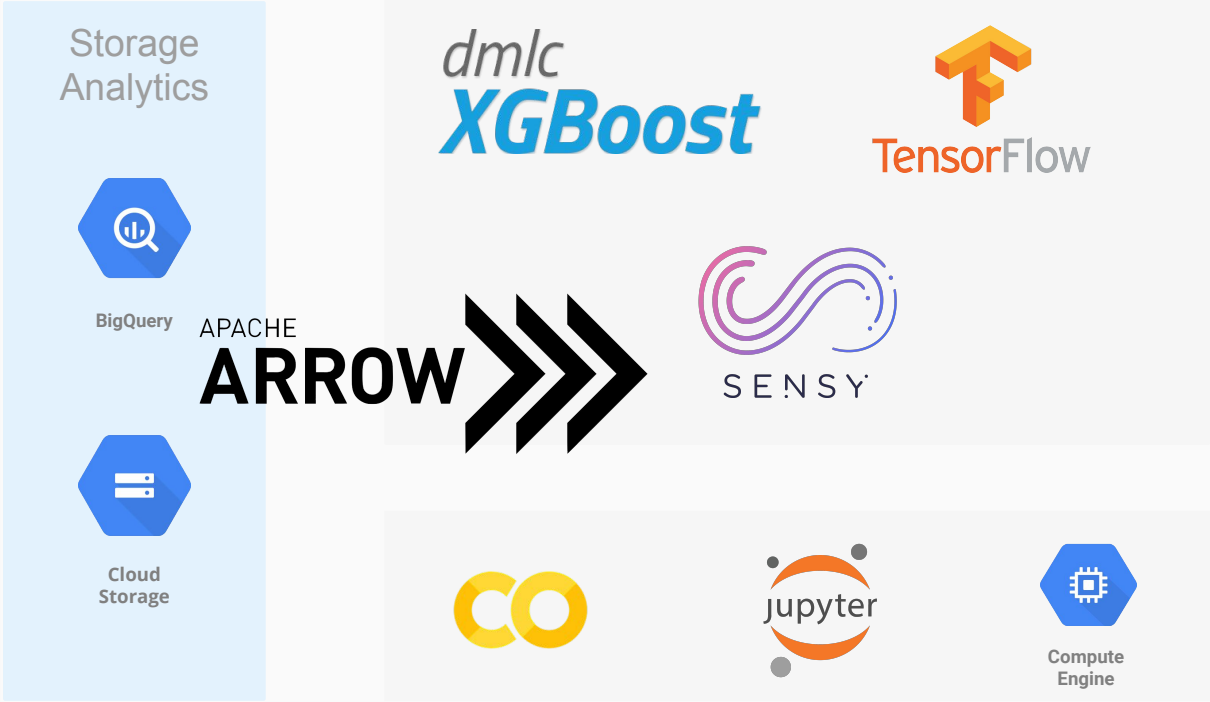
Apache Arrowで解決したこと

1. BQ Storage APIとの組み合わせによる転送の高速化
2. 互換性があるのでPandasが前提のライブラリでも利用可
3. ベクトルの保存と復元が簡単に

PoCでの課題

1. 数百GB ~ 数TBのデータ量
2. Pandasの利用
3. 密ベクトル

これらが解決され
試行錯誤が高速化



ケーススタディ: Taxi Trips

- Apache Arrow + BigQuery Storage API + TensorFlowでどのくらい高速化したか
- 検証に使ったColabノートブックはこちら
- データサイズ: 22.89GB
- レコード数:9419696件
- 次元数: 11

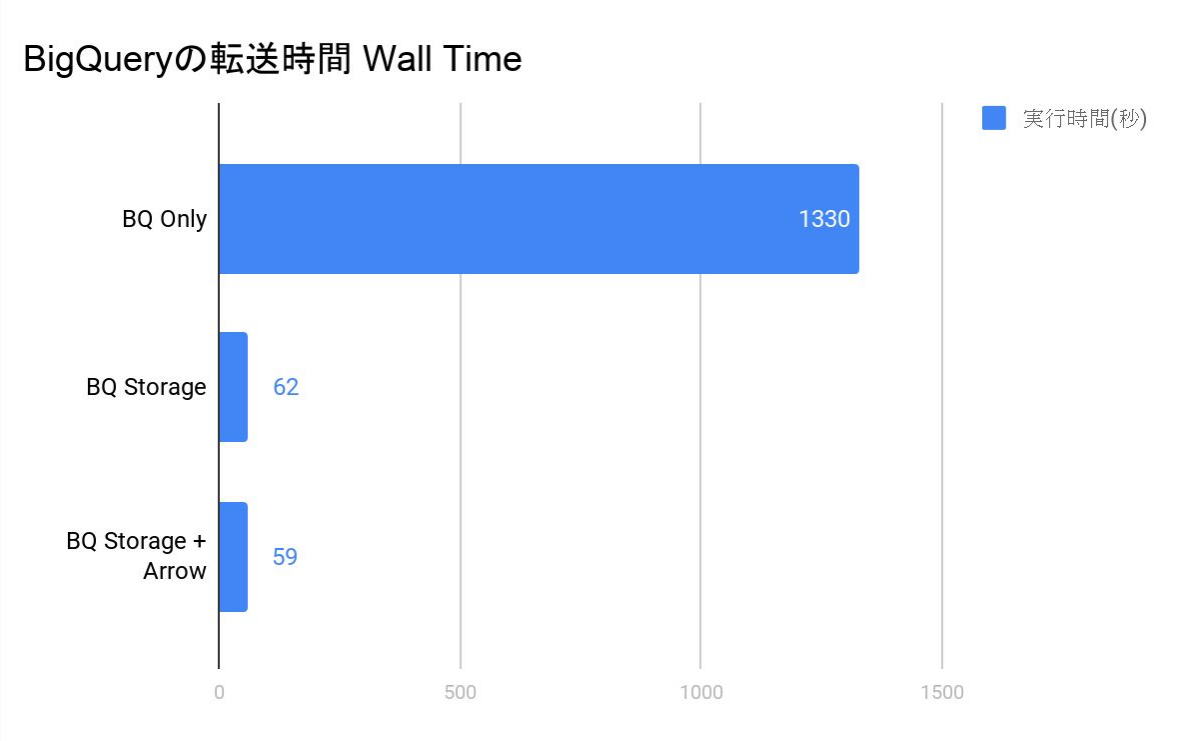
データ転送について

BQ + Arrowの転送

1. BigQuery Clientから転送
2. BigQuery Storage API
3. BigQuery Storage API
   + Apache Arrow

計測はwall timeによる計測

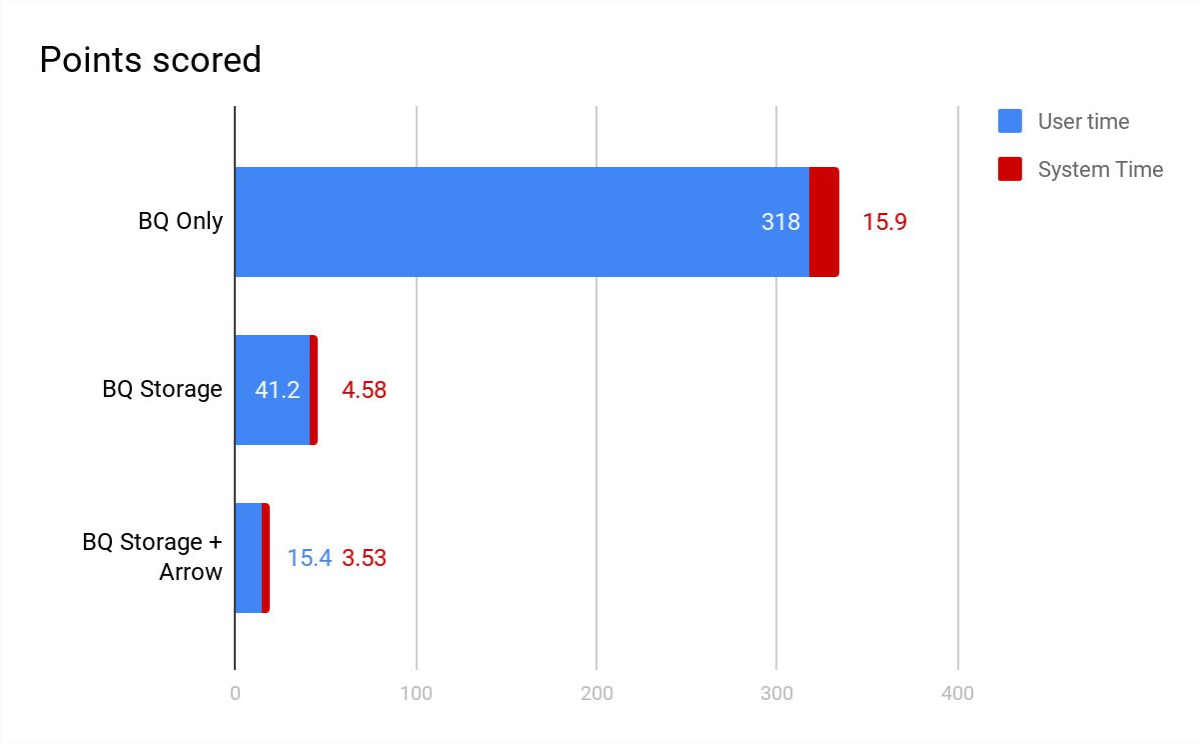
(ノートブックのセル実行終了時間)



BQ + Arrowの転送

1. BigQuery Clientから直接転送
2. BigQuery Storage API
3. BigQuery Storage API
   + Apache Arrow

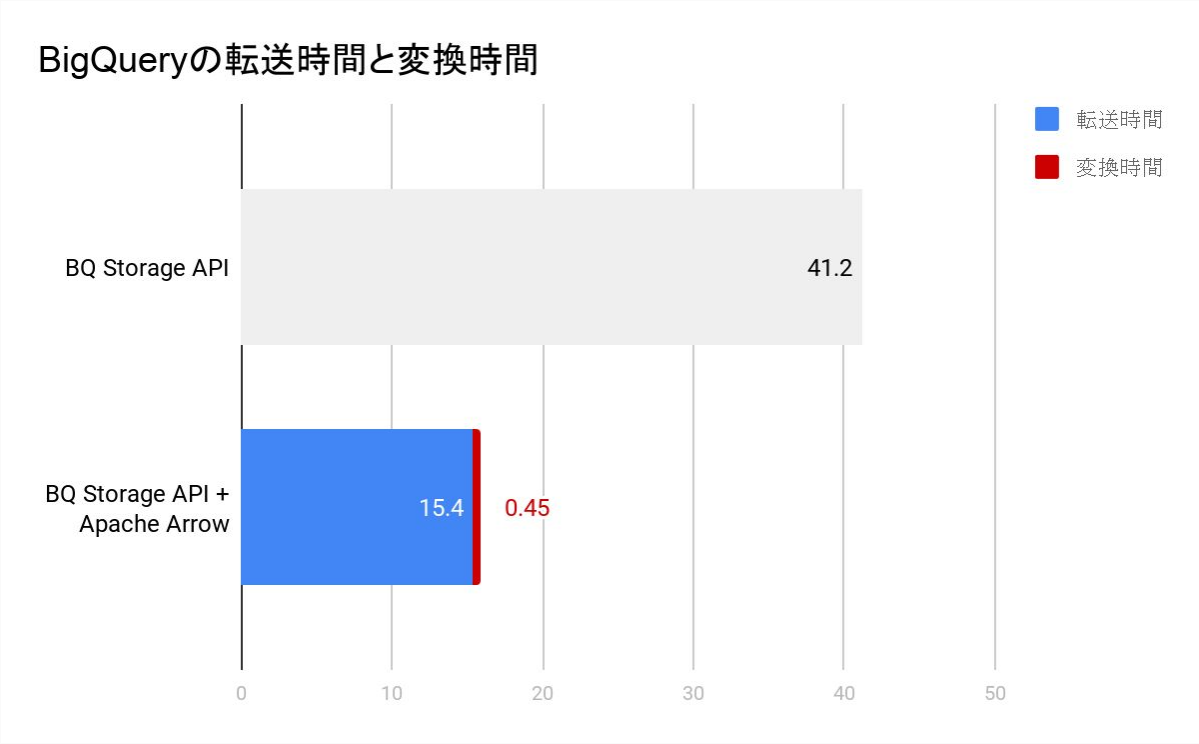
計測はUser Time + Sys Time



BQ Storage API転送 Pandas変換時間

1. BigQuery Storage API
2. BigQuery Storage API + Apache Arrow

計測はUser Time



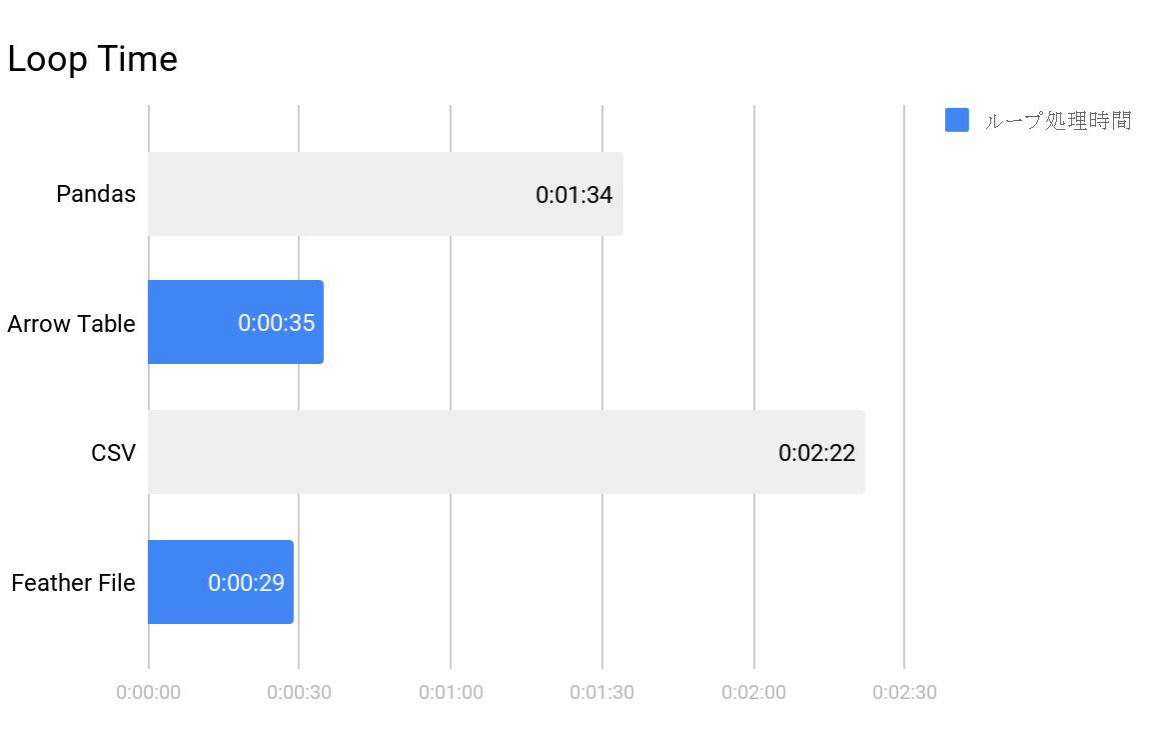
学習について

ループ時間計測

1. メモリに展開された状態のもの
   a. Pandas
   b. Arrow Table(RecordBatch)
2. Serializeされたもの
   a. CSV
   b. Feather

Walltime計測
TFRecordについては除外

tf.data.Dataset +
Keras.Model.compile
Keras.model.fit()

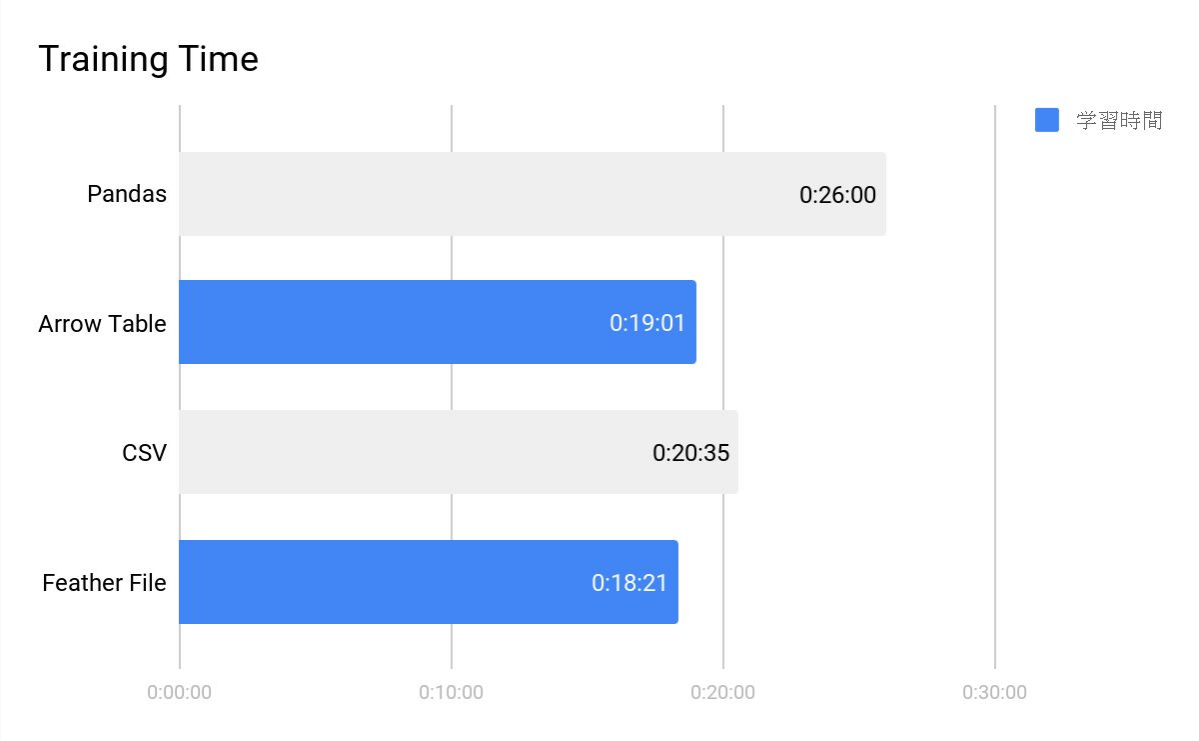


学習時間(GPU)

1. メモリに展開された状態のもの
   a. Pandas
   b. Arrow Table(RecordBatch)
2. Serializeされたもの
   a. CSV
   b. Feather

Wall Time計測
TFRecordについては除外

tf.data.Dataset +
Keras.Model.compile
Keras.model.fit



Apache Arrow
TensorFlow
BigQuery Storage API
早い・安い・すごい

ありがとうございました