Data Management for the World Wealth & Income Database

Christos Giatsidis, Antonis Skandalis, Konstantinos Skianis, Michalis Vazirgiannis
Facundo Alvaredo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, Gabriel Zucman
Ben Grillet, François Prosper, Brice Terdjman, Anthony Veyssière

DaSciM
Data Science and Mining Team
École Polytechnique

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

Project Overview

"The World Wealth and Income Database aims to provide open and convenient access to the most extensive available database on the historical evolution of the world distribution of income and wealth, both within countries and between countries."

Project's objectives

- Collect, process and publish income data at global level
- Provide the tools for multidimensional exploration and analysis of data
- Extract and demonstrate meaningful outcomes via effective visualisation tools

Technical challenges

- **BigData Approach:** Scalability in two dimensions:
  - Able to serve to patterns of user traffic that can reach $\sim 10^5$ hits/ hour
  - Able to deal with increasing amounts of time series data $\sim 10^9$ Samples
- Multi-attribute time series: Users query the database based on:
  - region of interest
  - period of interest in years
  - desired concept(s) (e.g. gdp per capita, wealth, income)
  - other attributes: population type, age, percentile
- Attribute and time based queries

Current Database Status (Apr. 2017)

Currently the database holds:

- 319 geographical regions (countries, continents, states)
- 150 combinations of attributes for each region
- data for around 50 years on average per region

http://www.wid.world

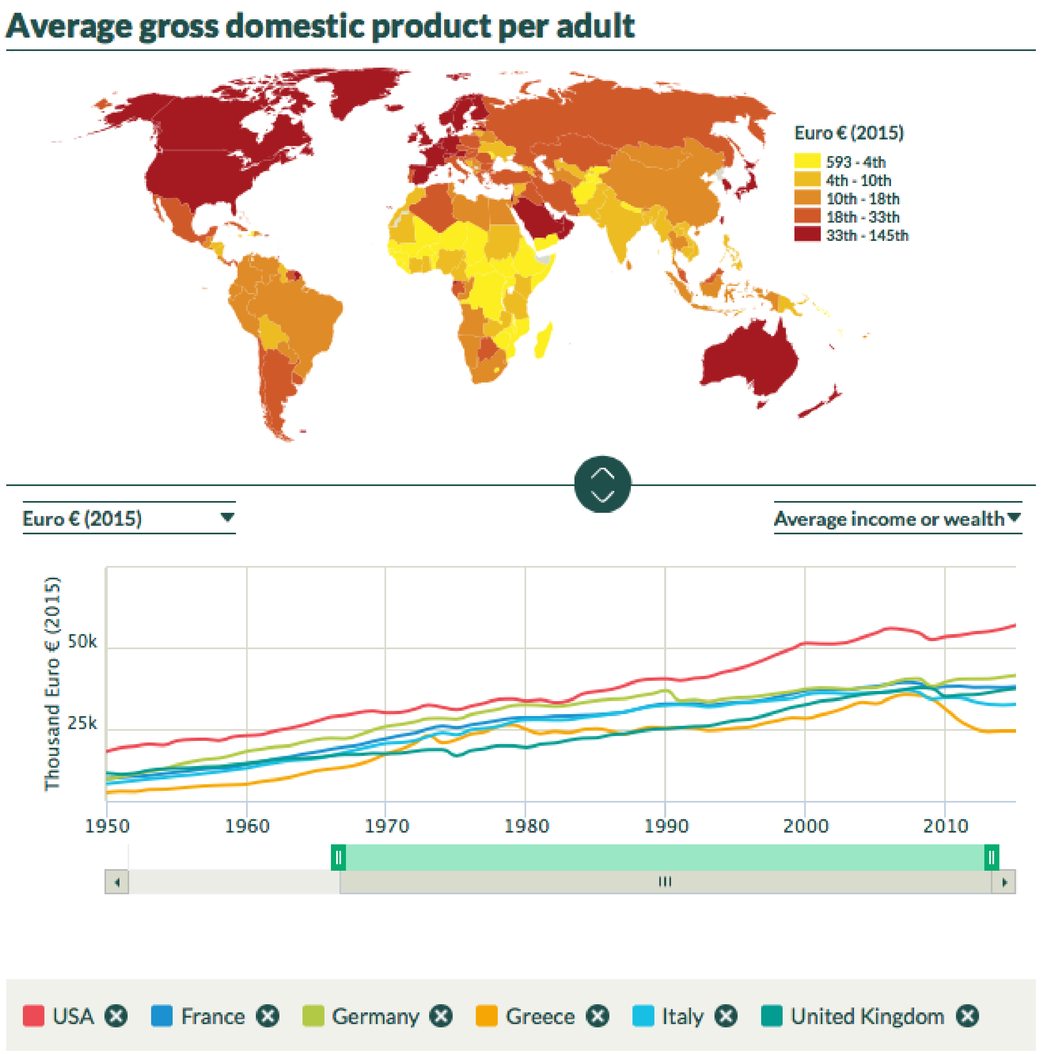


Figure: Data visualisation example at `http://www.wid.world`

Design

Key points

↪ **Cloud Implementation:** The project is hosted on Amazon

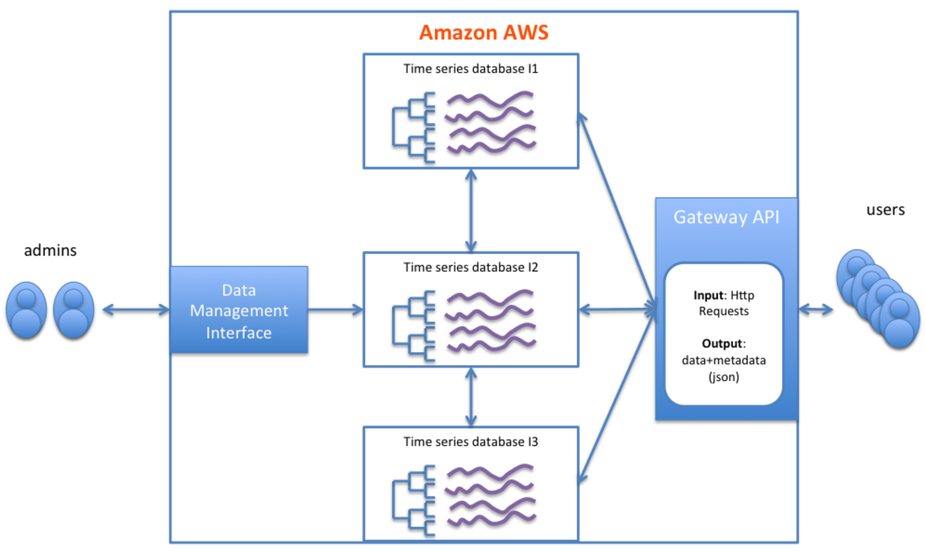
↪ **Web API data access:** Facilitating data sharing though an accessible API

↪ **Relational database implementation:** Achieving high performance queries by using state of the art indexing methods

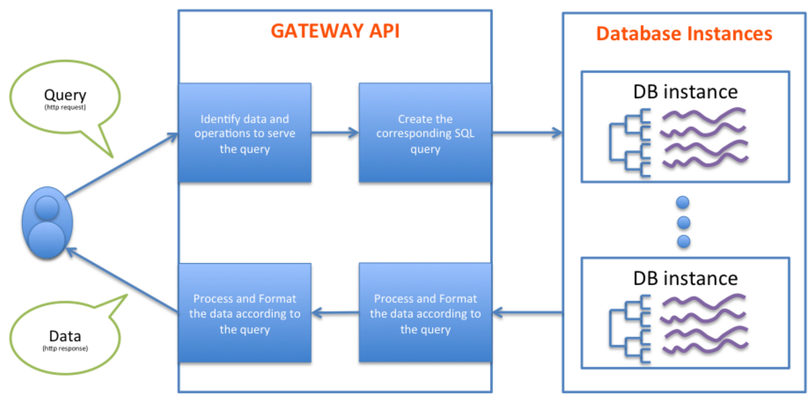
↪ **Scalability:** The database is designed in order to handle the increasing amounts of data

↪ **Data Management Interface:** A user-friendly web tool that enables the database maintenance

Architecture



Information Flow



Database

- Stores time series and hierarchically organized metadata
- Two layer approach:
  - Standard columns for *selective* attributes; each row corresponds to a unique combination
  - JSON format for other attributes and time series values
- Adaptive data schema:
  - possible addition/deletion/aggregation of features
  - complex feature hierarchy management

API

- Receives an HTTP request with the *query*
- The query is composed by data (country, concept, period etc) and operations (aggregation, conversion)
- Receives the corresponding data from the database by typical SQL queries on one of the available database instances
- The desired operations are applied on the time series
- Returns results in a JSON Format

Implementation Technologies

- Amazon GATEWAY API
- Database: Multi-instance Amazon RDS (PostgreSQL)
- Data Management Interface: Django, a python Web Framework
- Operations are implemented mostly in Numpy and Pandas Python Libraries

Credits

ECOLE D'ECONOMIE DE PARIS
PARIS SCHOOL OF ECONOMICS
World Inequality Lab @ PSE
(Project Leader and Owner)

THE PROJECT IS A JOINT COLLABORATION BETWEEN:

DaSciM
Data Science and Mining Team @ Ecole Polytechnique
(Database Architecture and Optimization)

WEDODATA
Web Design & Visualisation

dascim@lix.polytechnique.fr