A Tensor Based Data Model for Polystore

Presented by Samyak Ahuja and Mayank Kharbanda

Prerequisites

Data Warehouse

Polystore

Tensors

Imagination*

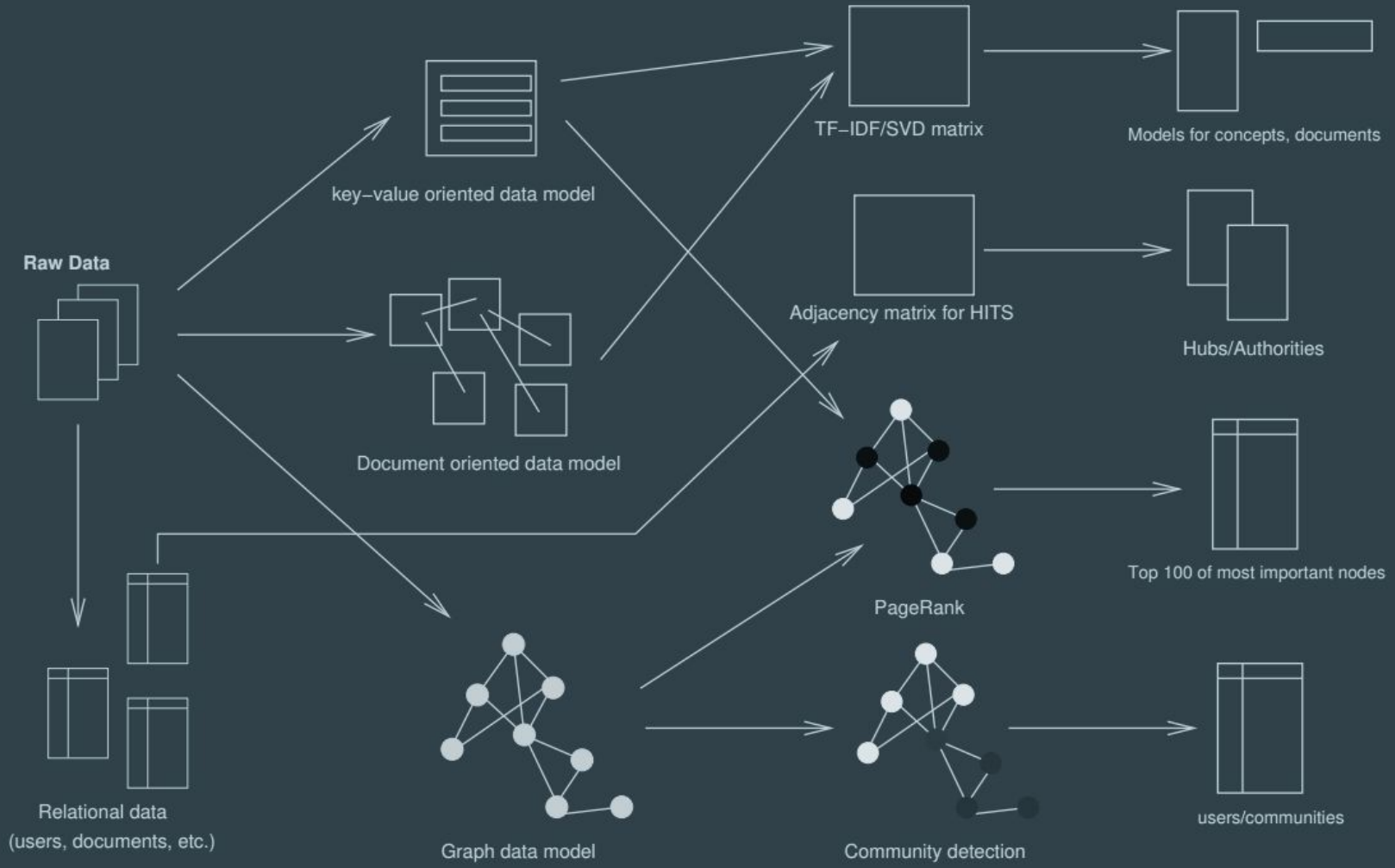
Data Warehouse

Application to Social Network Data

Data from Social Networks like Twitter is becoming increasingly relevant to Social Sciences and have become a good playground for Data Scientists to work on.

What a Data Scientist needs is

1. Gain Control Over the Data
2. Build efficient algorithms for analysis


Raw Data
key–value oriented data model
Document oriented data model
TF–IDF/SVD matrix
Models for concepts, documents
Adjacency matrix for HITS
Hubs/Authorities
PageRank
Top 100 of most important nodes
Relational data
(users, documents, etc.)
Graph data model
Community detection
users/communities

Polystore

In recent years we have seen the convergence of two different fields, namely, High Performance Computing(HPC) and Databases to form a new field called Data Intensive HPC. One of the major concerns of Data Intensive HPC is to be able to quickly feed the Algorithm with the required data.

Drawbacks of Single Data Model

Transforming various data into a single model may have a significant impact on performance of queries but also on capabilities to apply different algorithms.

Taking inspiration from distributed databases we try to create a federation of specialized storage systems with different models.

Now each algorithm might be using a different data model so to accommodate that we build a multi-paradigm storage system called Polystore.

In such systems data can be partitioned and stored in a model that best fits the algorithm required for analysis.

Abstract

We show how the mathematical object tensor can be used to build a multi-paradigm model for the storage of social data in data warehouses.

Our approach allows to link different storage systems (polystore) and limits the impact of ETL tools performing model transformations required to feed different analysis algorithms.

The proposed model allows to reach the logical independence between data and programs implementing analysis algorithms.

Now that we have decoupled the data from the analysis algorithms using polystores, we need to work on simplifying model transformations.

The research paper contributes towards the definition of a data model based on tensors that does exactly this.

Here tensor is rather a structure of exchange among processes of complex workflow rather than a model to represent real data.

Tensors