A Scalable Data Transformation Platform

– The Example of Data Anonymization

Xiufeng Liu, *Postdoc.*

xiuli@dtu.dk

Department of Engineering Management, DTU

26 May 2015

Table of Contents

CITIES

DTU

Introduction

- In data warehousing, up to 80% of the time is spent on data pre-processing [1], including data extraction, transformation and loading;
- It is challenging for big data transformation and analytics;
- Diverse data warehousing tools in the market incorporate big data technologies for handling scalable data sets, but most of them lack of flexibility;

CITIES

DTU

Introduction

- In data warehousing, up to 80% of the time is spent on data pre-processing [1], including data extraction, transformation and loading;
- It is challenging for big data transformation and analytics;
- Diverse data warehousing tools in the market incorporate big data technologies for handling scalable data sets, but most of them lack of flexibility;

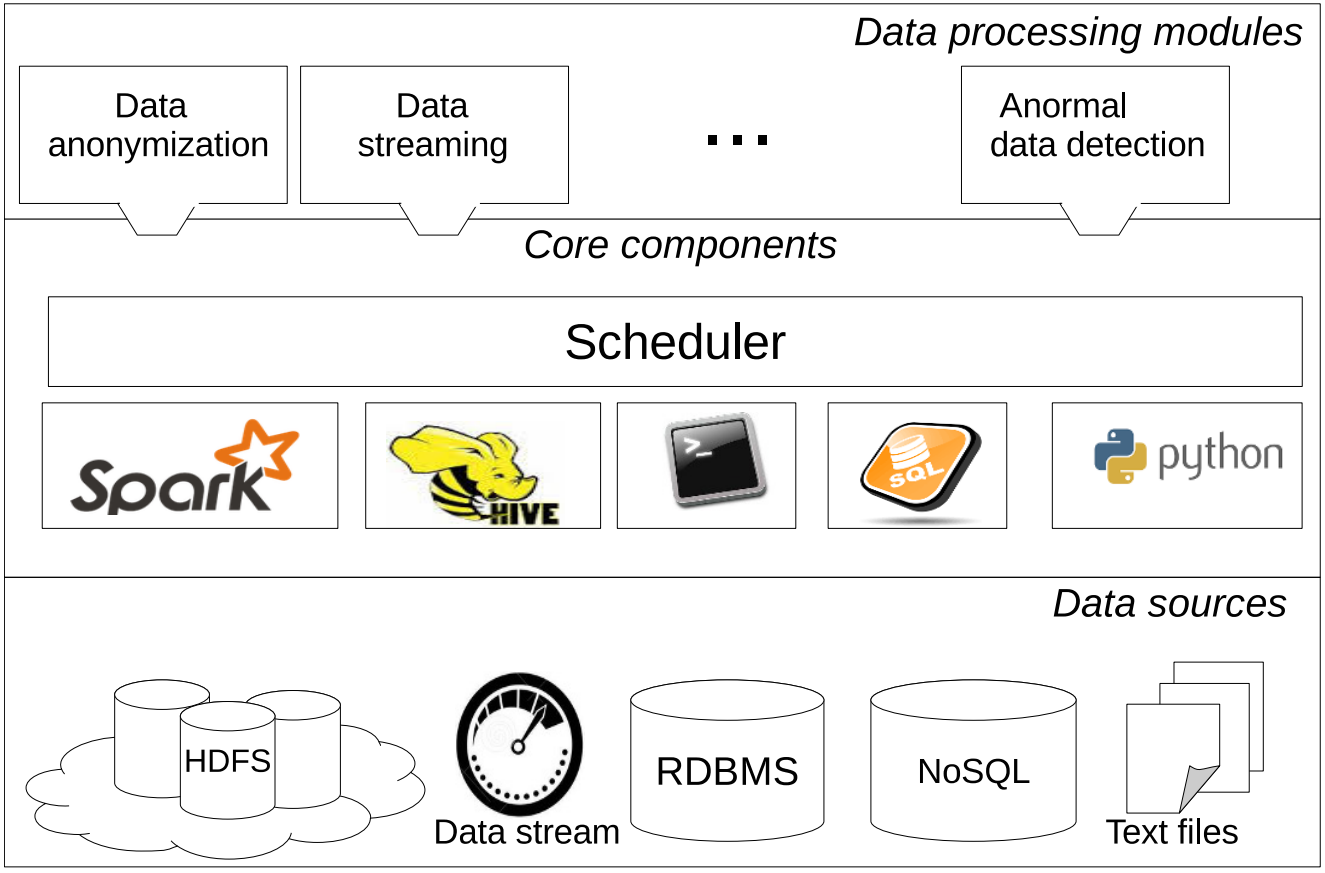
Therefore, we intend to implement a scalable platform for big data transformation and analytics; and for research and production.

CITIES

DTU

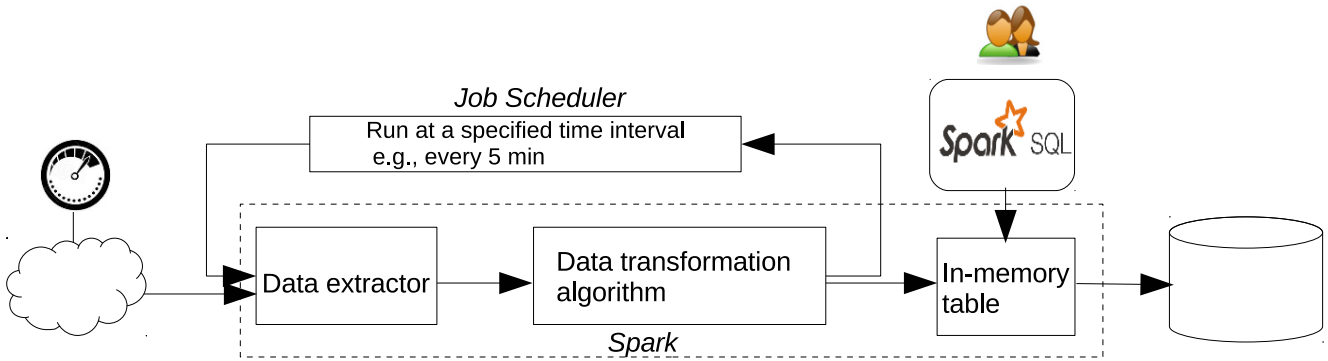
System Architecture

- The building blocks of the system



Realtime & Batch Processing

- Realtime stream processing
  - Spark Streaming (*in-memory based distributed computing framework*)



- Batch processing
  - Hive (*built on top of Hadoop MapReduce framework*)
- Job scheduling strategies
  - Deterministic – *For jobs not running in a cluster*
  - Undeterministic – *For jobs running in a cluster*

CITIES

DTU

A Transformation Example – Data Anonymization (Cont.)

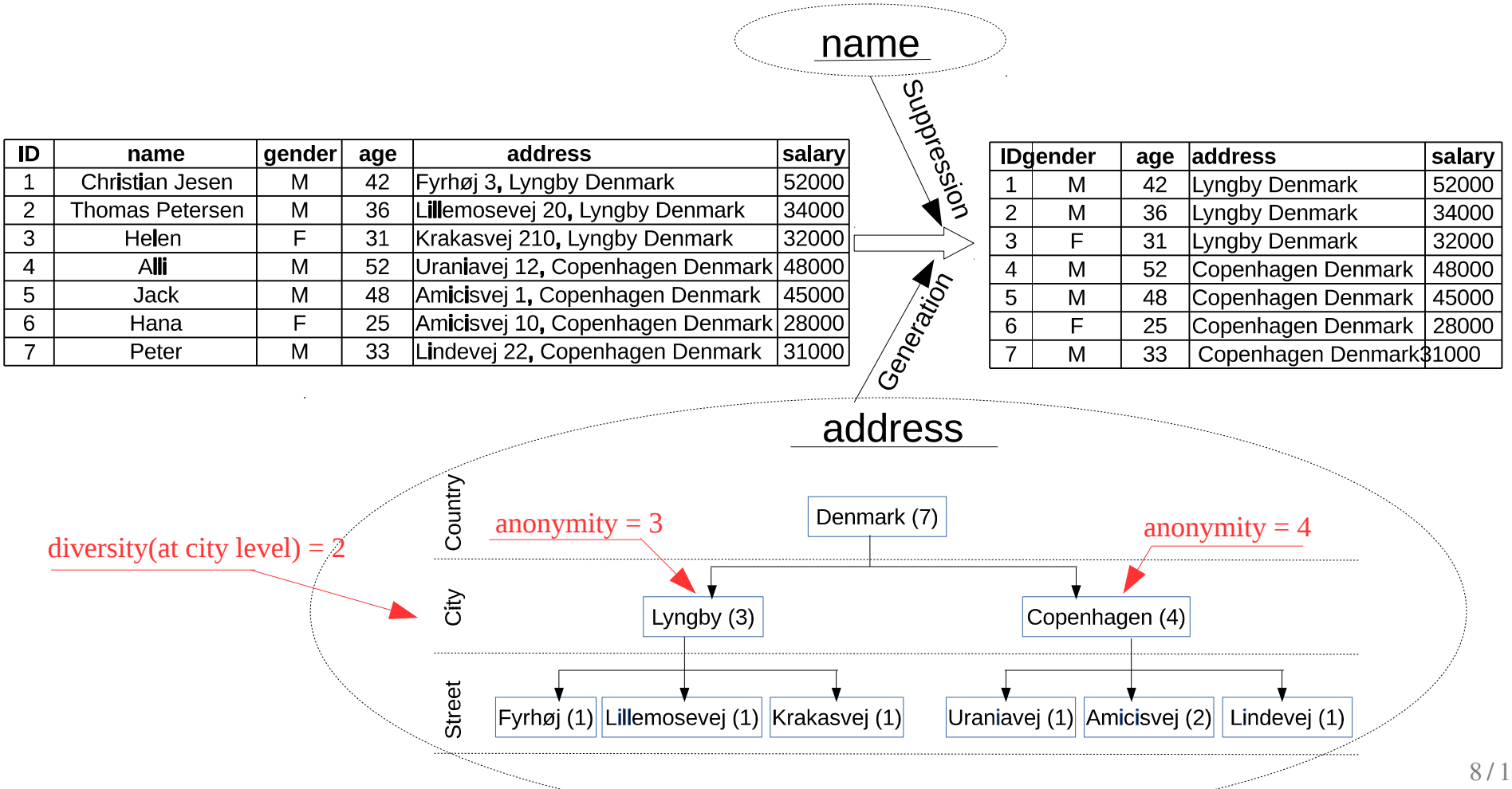
- Anonymization Process:
  - *Data Source* $\rightarrow$ *Anonymize* $\rightarrow$ *Data Publishing & Sharing*
- Anonymization for two types of data:
  - Smart meter data:
    - Separate storing meter data from social-economic data, and use foreign-key referencing between tables
    - Aggregate data if possible
  - Social-economic data, e.g., customer information
    - Anonymize methods: *generalization* and *suppression*
    - Metric: *k-anonymity* and *l-diversity*
    - Balancing between *anonymity level* and *information loss*

CITIES

DTU

A Transformation Example – Data Anonymization (Cont.)

- Anonymize quasi-attribute values, i.e., *name* and *address*. If the requirement of anonymizing *address* attribute values is *k-anonymity*≥ 2 and *l-diversity*≥ 2, then generalize to *city* level will suffice.



ID	name	gender	age	address	salary
1	Christian Jesen	M	42	Fyrhøj 3, Lyngby Denmark	52000
2	Thomas Petersen	M	36	Lillemosevej 20, Lyngby Denmark	34000
3	Helen	F	31	Krakasvej 210, Lyngby Denmark	32000
4	Alli	M	52	Uraniavej 12, Copenhagen Denmark	48000
5	Jack	M	48	Amicisvej 1, Copenhagen Denmark	45000
6	Hana	F	25	Amicisvej 10, Copenhagen Denmark	28000
7	Peter	M	33	Lindevej 22, Copenhagen Denmark	31000

ID	gender	age	address	salary
1	M	42	Lyngby Denmark	52000
2	M	36	Lyngby Denmark	34000
3	F	31	Lyngby Denmark	32000
4	M	52	Copenhagen Denmark	48000
5	M	48	Copenhagen Denmark	45000
6	F	25	Copenhagen Denmark	28000
7	M	33	Copenhagen Denmark	31000

A Transformation Example – Data Anonymization

- **Use Case:** We have a remote server containing customer data. We want to extract, anonymize, and publish the data.
- **The implementation:**
  1. Extract the raw customer data from remote server to staging database at local server, e.g., by SCP;
  2. Define the anonymization rules for suppressing/generalizing quasi-attribute values;
  3. Verify the results (by the online analytics tool)
  4. Load the anonymized data into database for publishing.
  5. Housekeeping, i.e., delete the data from staging database, and send email for notification, etc.

CITIES

DTU

References

C. Thomsen and T. B. Pedersen. pygrametl: A powerful programming framework for Extract-Transform-Load Programmers. In Proc. of DOLAP, pp. 49-56, 2011.

CITIES

DTU