



POLITECNICO
MILANO 1863



Performance and Cost Evaluation of Privacy Preservation in Big Data Systems

4 December 2017

Candidate: Safia Kalwar
Mentor: Prof. Danilo Ardagna

DICE Horizon 2020 Project
Grant Agreement no. 644869
<http://www.dice-h2020.eu>



Funded by the Horizon 2020
Framework Programme of the European Union

OUTLINE



- Problem Statement
- Objective
- Methodology
- Conclusions & Future Work

Problem Statement



Big data burst upon in the first decade of the 21st century

IBM claims 90% of today's stored data was generated in just last two years.



Security and privacy issues are magnified by velocity, volume and variety of big data.



Large scale cloud infrastructure, diversity of software platforms, spread across large networks of computers, increase surface attack of the entire system.

Research Objective



- To investigate the performance of multiple privacy protection mechanisms.
- To investigate the cost impact of multiple privacy protection mechanisms.



Privacy Mechanism



- **Data Anonymization:** de-identification, consists of techniques that can be applied to prohibit the recovery of individual information.

Sensitive attributes	Quasi-identifiers		attributes
Name	Age	Education	Disease
Alice	19	10 th	Diabetes
Jean	19	9 th	Cancer
Ines	27	9 th	Flu
David	30	9 th	Flu
Bob	23	11 th	Cancer
Dupont	23	11 th	Cancer

Perturbation

Name	Age	Education	Disease
Clarie	19	10 th	Diabetes
John	19	9 th	Cancer
Agnes	27	9 th	Flu
Peter	30	9 th	Flu
Ana	23	11 th	Cancer
David	23	11 th	Cancer

Privacy Mechanism



- **Encryption:** The technique uses cryptographic schemes based on public or symmetric key to replace sensitive data.

Sensitive attributes	Quasi-identifiers		attributes
Name	Age	Education	Disease
Alice	19	10 th	Diabetes
Jean	19	9 th	Cancer
Ines	27	9 th	Flu
David	30	9 th	Flu
Bob	23	11 th	Cancer
Dupont	23	11 th	Cancer

Encryption

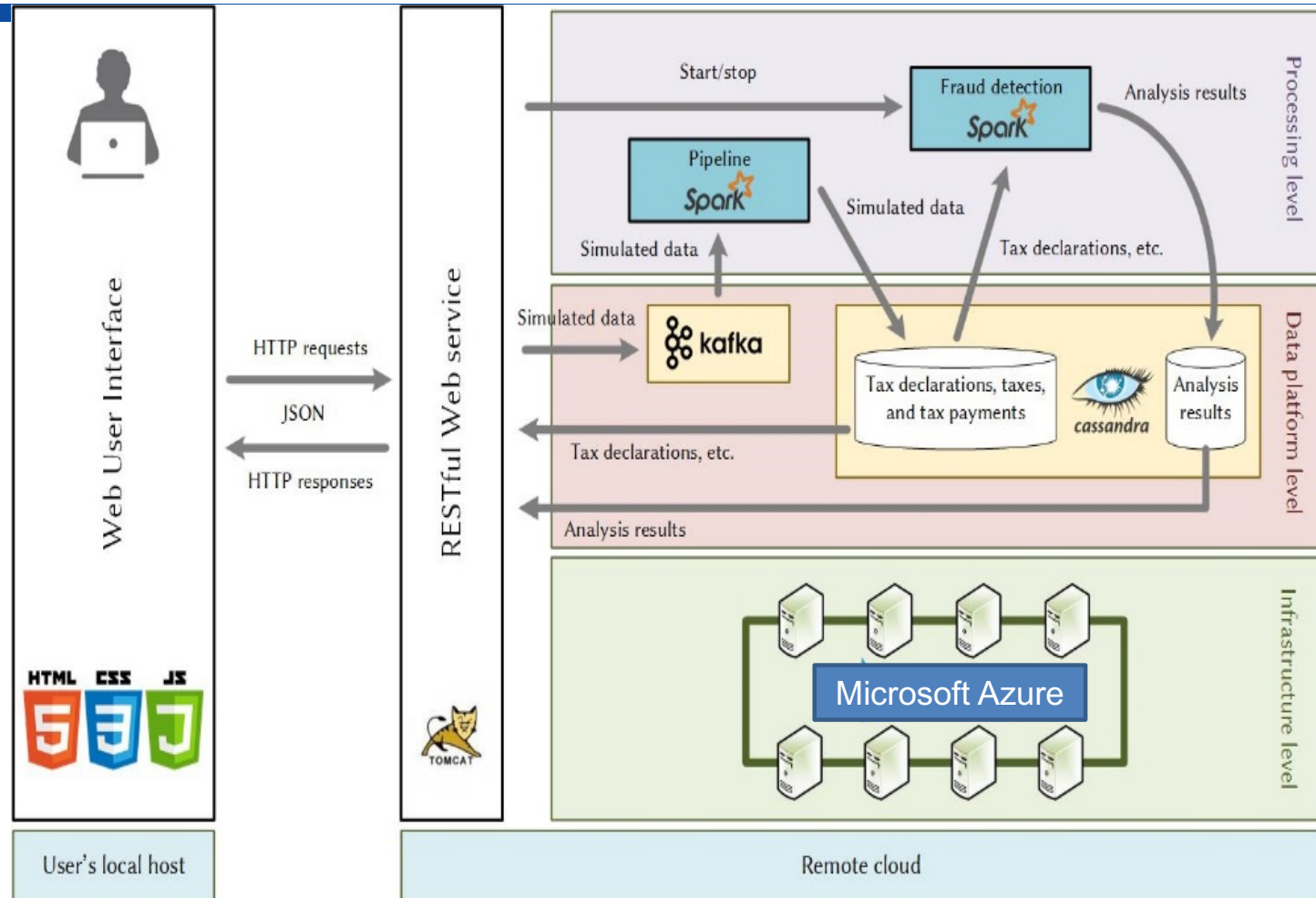
Sensitive attributes	Quasi-identifiers		attributes
	Age	Education	Disease
85ul			
38ur	19	10 th	Diabetes
9iyi	19	9 th	Cancer
455r	27	9 th	Flu
454t	30	9 th	Flu
878i	23	11 th	Cancer
8u7y	223	eur7	bra5

We used AES 128 and 256 bit encryption algorithms.

NETF case study



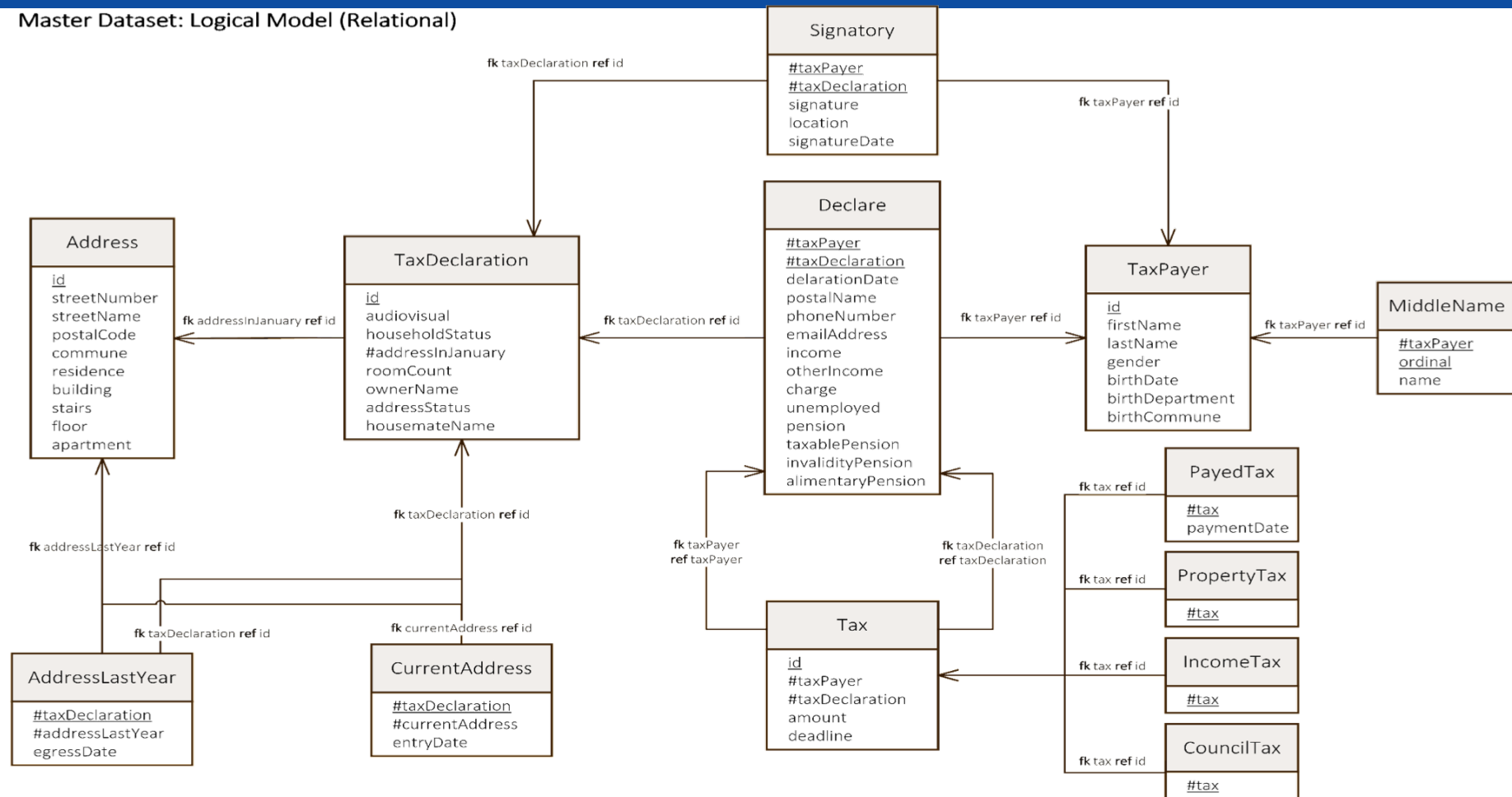
- Big data application for tax fraud detection.
- Goal: build a model of “fraudulent conduct” from automatic operations on existing tax data files (e.g. business creation or bank accounts), which represent millions of data points.



Relational database model



Master Dataset: Logical Model (Relational)



Fraud Indicator



- Here on the top 3 queries are introduced and then the proposed privacy techniques those are masking and encryption.

Query 1:

- Compares the income of current year to previous year.
- Detects the fraudulent by setting criteria (criteria could vary Example 20%).

Query 5:

- Contains all the information needed to know every individual income and to justify the amount of tax to be paid.

Query 7:

- Shows the information about the location and the date of the payment.

Derived queries



- **Query 3:**
 - Derived from Query 1.
 - Designed for reading masked data.
 - Measure the overhead on the system due to the additional join with the dictionary table.

- **Query 6 and 8:**
 - Replicate queries 5 and 7 respectively with additional join.

Optimization Tool (D-SPACE4Cloud)



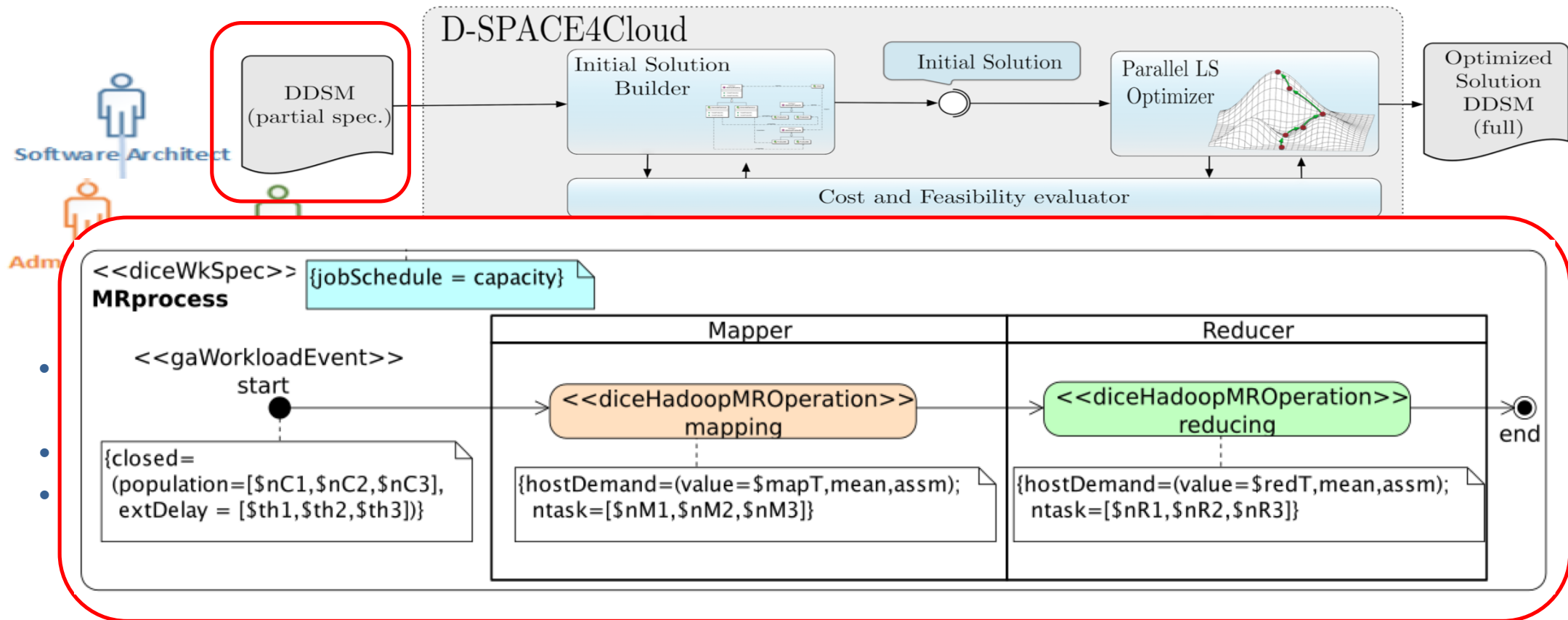
The problem:

- Minimize costs and suggests the optimal deployment architecture which provides performance guarantees:
 - Given a deadline D .
 - Select the VMs type and their number, s.t. $E[R] \leq D$.
- Novel tool, design space exploration has been increasingly sought in traditional multi-tier applications, but not in the design of Data intensive Applications.

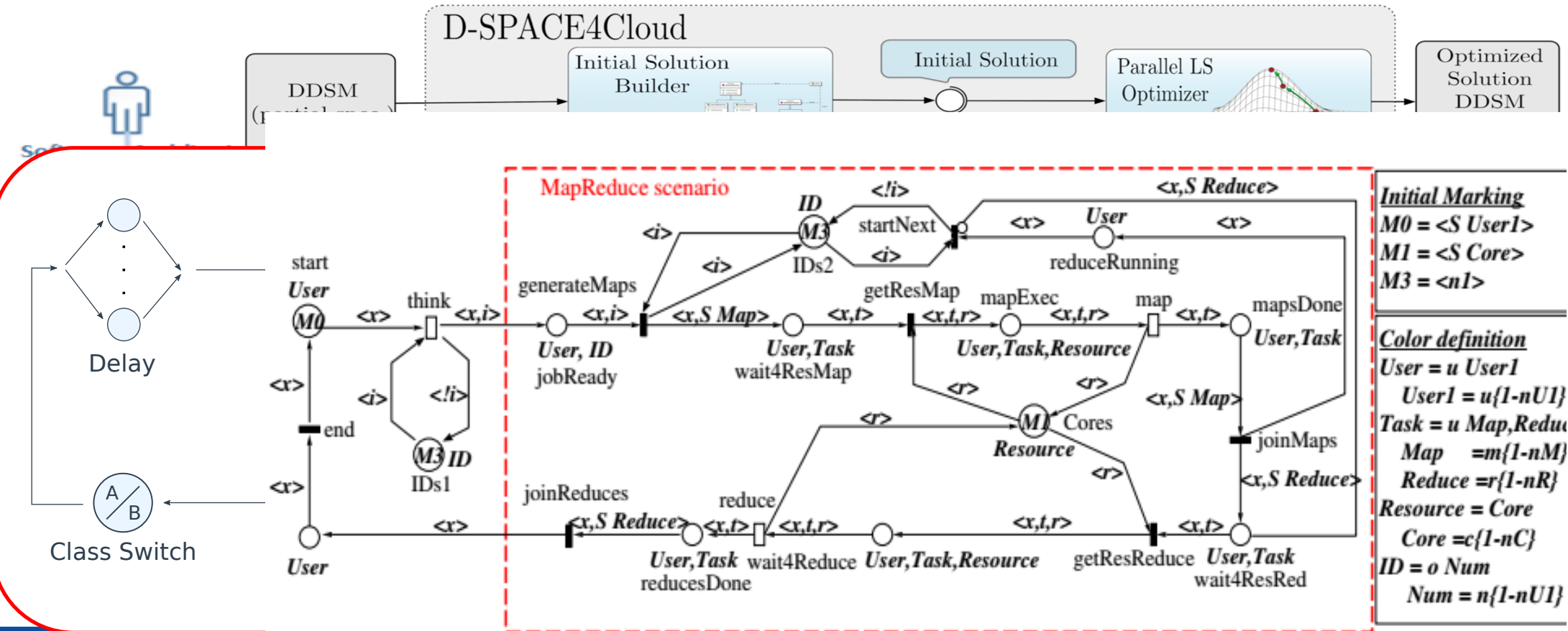
What does the tool do?

- Automatic analysis of multiple candidate alternative configurations.
- Relies on simulation to predict application performance.

D-SPACE4Cloud Architecture



D-SPACE4Cloud Architecture



Configuration Setup



■ Performance Analysis

■ Virtual Machines

- D4V2
- D12V2

■ Configuration parameters

- Varying scale
 - Driver memory, Executor memory, Executor cores
- Virtual machines: 3-13
- Executors: 4
- Cores: 2

■ Cost Analysis

■ Virtual Machines

- D4V2
- D12V2

- Varying deadline
- Specific to the query

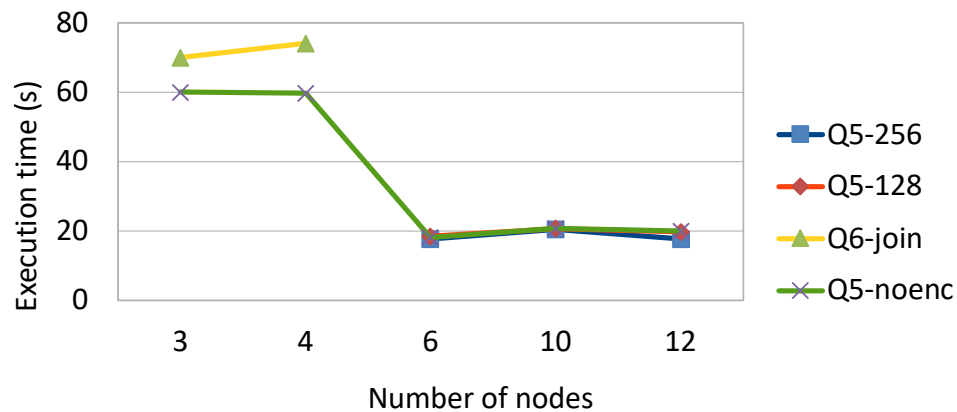
Parameter	Value	Units of Measurements
	D12v2	
Cores	4	---
Memory	28	[GB]
Cost	0.25	[€/h]
	D4v2	
Cores	8	---
Memory	28	[GB]
Cost	0.25	[€/h]

DATASET GENERATION

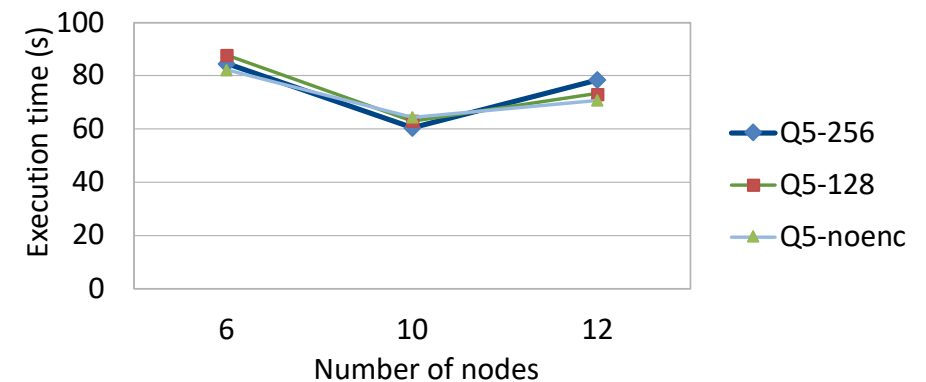


- For performance and cost analysis many datasets has been generated varying their size (number of records in Tax Declaration table).
- **NETF**
 - From 1000 up to 30,000,000
 - 13 different datasets have been generated

Performance Evaluation analysis NETF

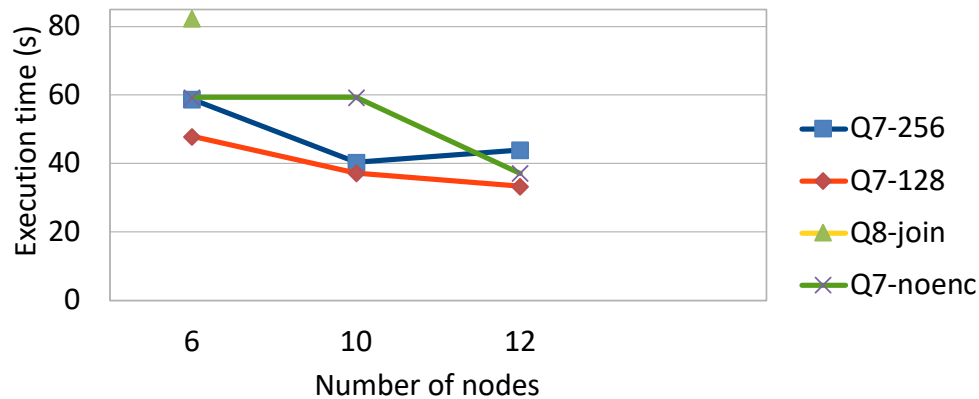


Performance degradation analysis of Query 5,6 for 1.5 million entries data set

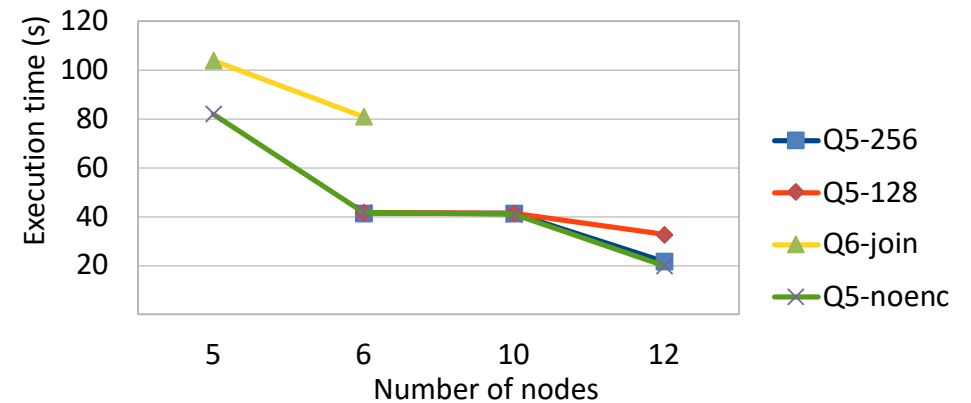


Performance degradation analysis of Query 5 for 30 million entries data set

Performance Evaluation analysis NETF

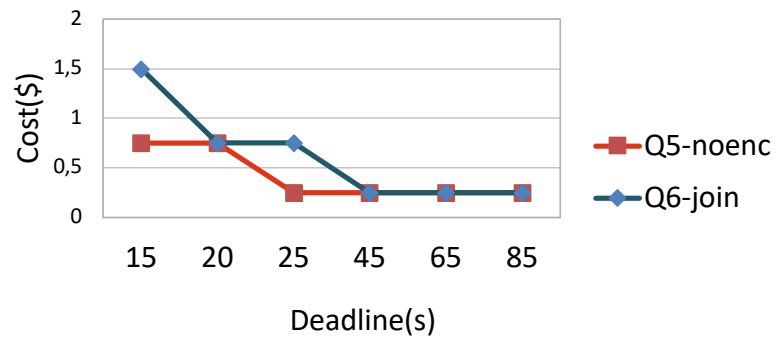


Performance degradation analysis of Query 7,8 for 10 million entries data set

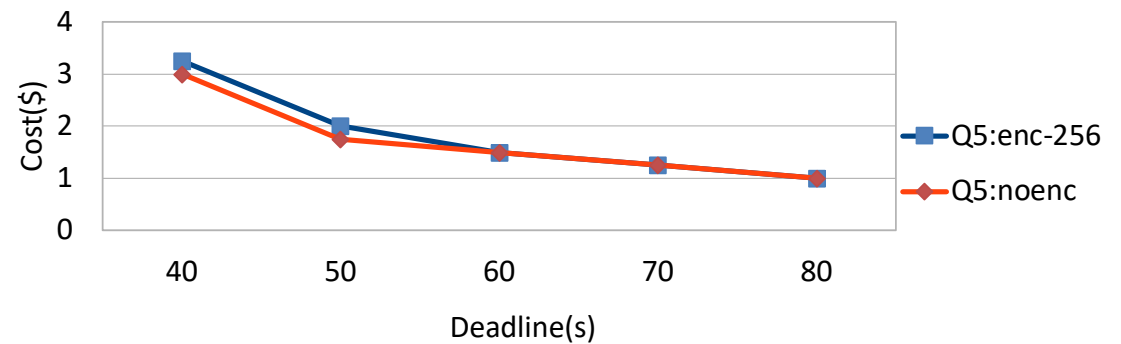


Performance degradation analysis of Query 5,6 for 13 million entries data set

Cost Evaluation analysis **NETF**

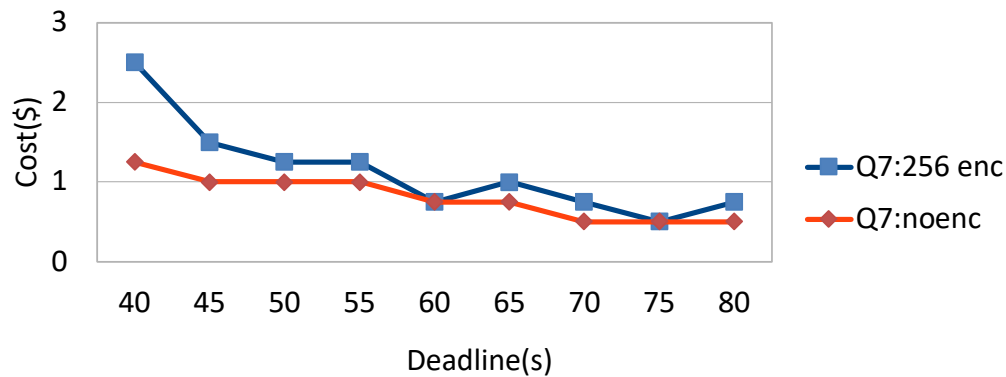


Cost Evaluation of Query 5,6 for 1.5 million entries data set.

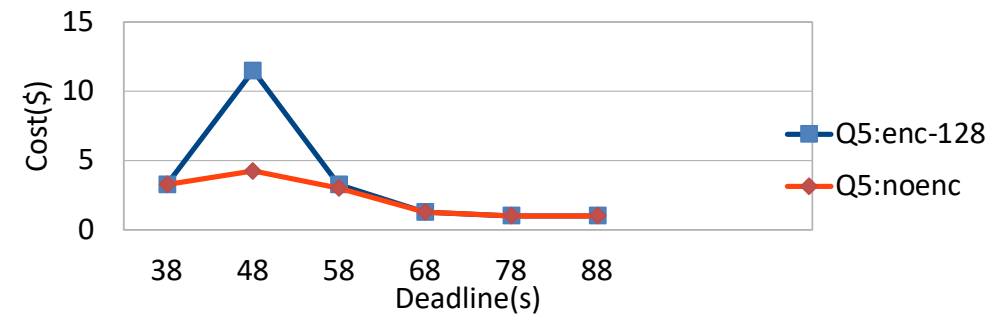


Cost Evaluation of Query 5 for 30 million entries data set.

Cost Evaluation analysis **NETF**



Cost Evaluation of Query 7 256 bit encrypted, unencrypted for 10 million entries data set.



Cost Evaluation of Query 5 128 bit encrypted, unencrypted for 13 million entries data set.

Results Summary



- Limited overhead
- Not always performance degradation introduces additional cost.

Benchmark		Performance degradation %	Cost overhead %
NETF	Masking	30	66
	Encryption	25	50
TPC-DS	Masking	10	10
	Encryption	30	25

Conclusions and Future work



- Performance and cost analyses of privacy mechanisms implementation in big data applications.
 - Performance overhead and cost impact are application specific.
 - Needs for tools to estimate privacy impact at design.
-
- Additional validation (e.g., hash on TPC-DS).
 - Automation of test generation and execution.



Thank you for your Time

Any Questions?
