Top Down Specialization for Apache Spark™

Why data privacy?

A city’s voter list was used to identify voters’ medical records

97% of voters were identified by only using ZIP Codes and birth dates

A New York Times reporter identified a woman by using her web searches

96% of Netflix subscribers were uniquely identified in 2006

Important Definitions

Quasi-Identifiers

Attributes that when combined together can identify an individual

Sensitive Attributes

Attributes that we are trying to conceal when datasets are released

Taxonomy Trees

Hierarchy of distinct values in a dataset

k-anonymity

- A dataset is called k-anonymous when for ever record there are at least $k\mathbf{-1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Master’s	Male	Orleans	$50,000
PhD	Male	Gloucester	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

k-anonymity

- A dataset is called k-anonymous when for ever record there are at least $k-\mathbf{1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Master’s	Male	Orleans	$50,000
PhD	Male	Gloucester	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

k-anonymity

- A dataset is called k-anonymous when for ever record there are at least $k\mathbf{-1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Master’s	Male	Orleans	$50,000
PhD	Male	Gloucester	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

k-anonymity

- A dataset is called k-anonymous when for every record there are at least $k - \mathbf{1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Master’s	Male	Orleans	$50,000
PhD	Male	Gloucester	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

k-anonymity

- A dataset is called k-anonymous when for every record there are at least $k-\mathbf{1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Master’s	Male	Orleans	$50,000
PhD	**Male**	**Gloucester**	**$100,000**
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

k-anonymity

- A dataset is called k-anonymous when for every record there are at least $k-\mathbf{1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Master’s	**Male**	**Orleans**	**$50,000**
PhD	Male	Gloucester	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

k-anonymity

- A dataset is called k-anonymous when for every record there are at least $k-\mathbf{1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Master’s	Male	Orleans	$50,000
PhD	Male	Gloucester	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

k-anonymity

- A dataset is called k-anonymous when for every record there are at least $k-\mathbf{1}$ records with the same quasi-identifier values

Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
~~Master’s~~ Graduate	Male	~~Orleans~~ Ottawa East	$50,000
~~PhD~~ Graduate	Male	~~Gloucester~~ Ottawa East	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

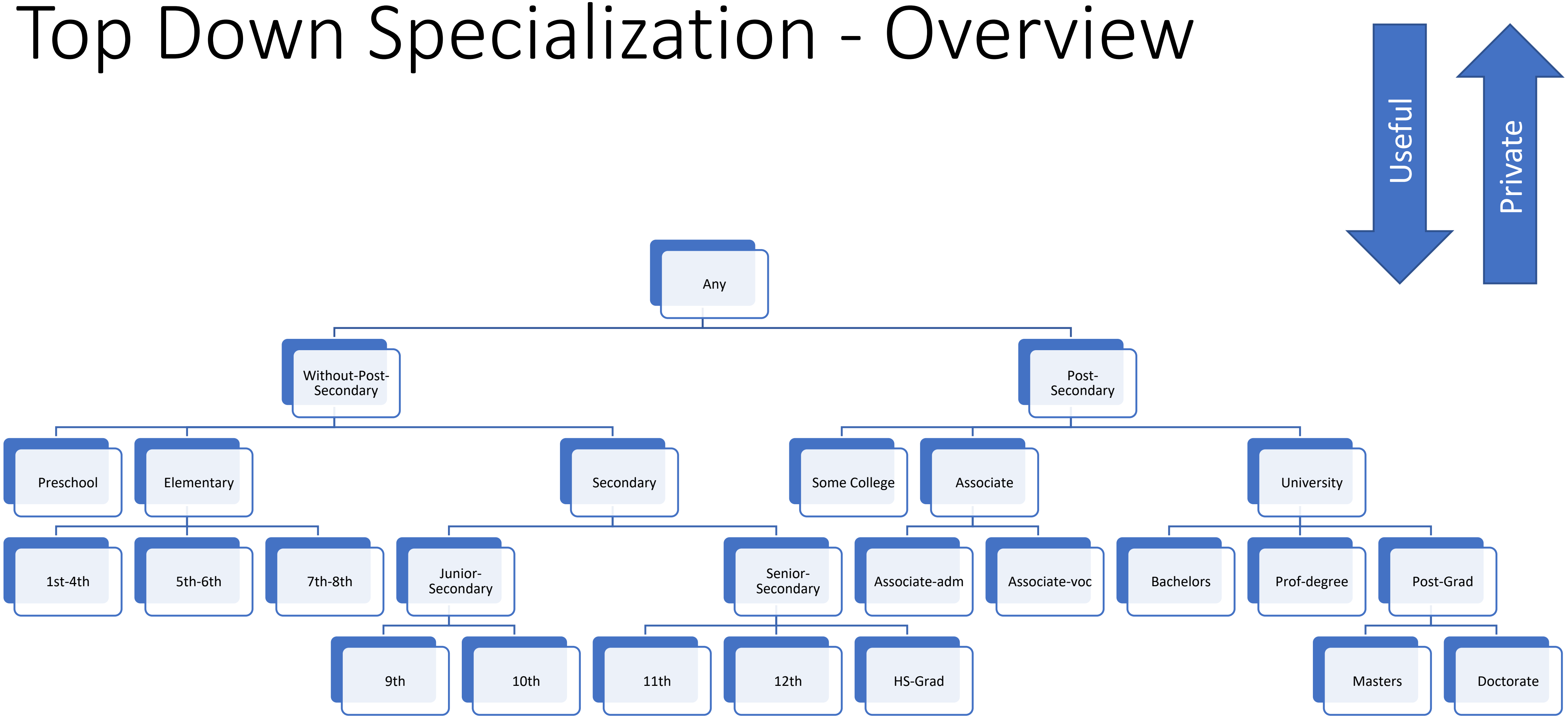
k-anonymity

- A dataset is called k-anonymous when for every record there are at least $k\mathbf{-1}$ records with the same quasi-identifier values

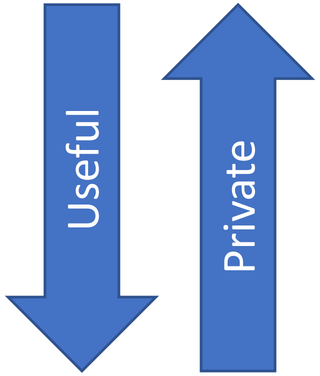
Education	Gender	City	Income
Grade 12	Female	Nepean	$65,000
Bachelor’s	Male	Ottawa	$50,000
Graduate	Male	Ottawa East	$50,000
Graduate	Male	Ottawa East	$100,000
Grade 12	Female	Nepean	$80,000
Associate	Female	Kanata	$90,000
Associate	Female	Kanata	$105,000
Bachelor’s	Male	Ottawa	$50,000

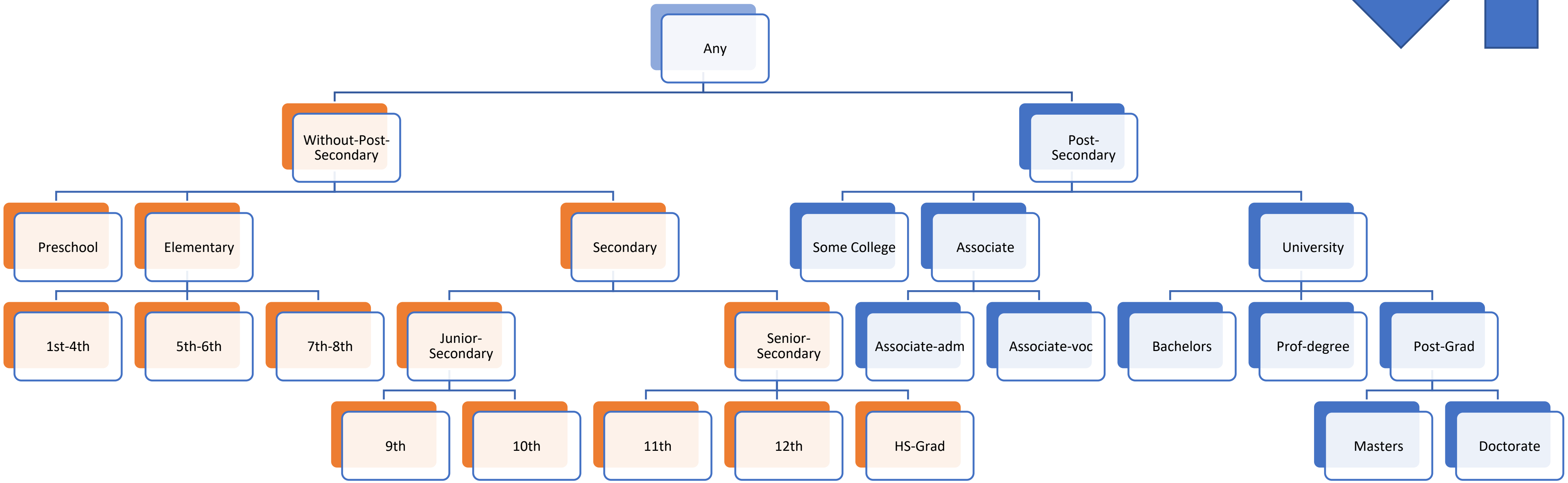
2-anonymous

Top Down Specialization - Overview



Top Down Specialization - Overview



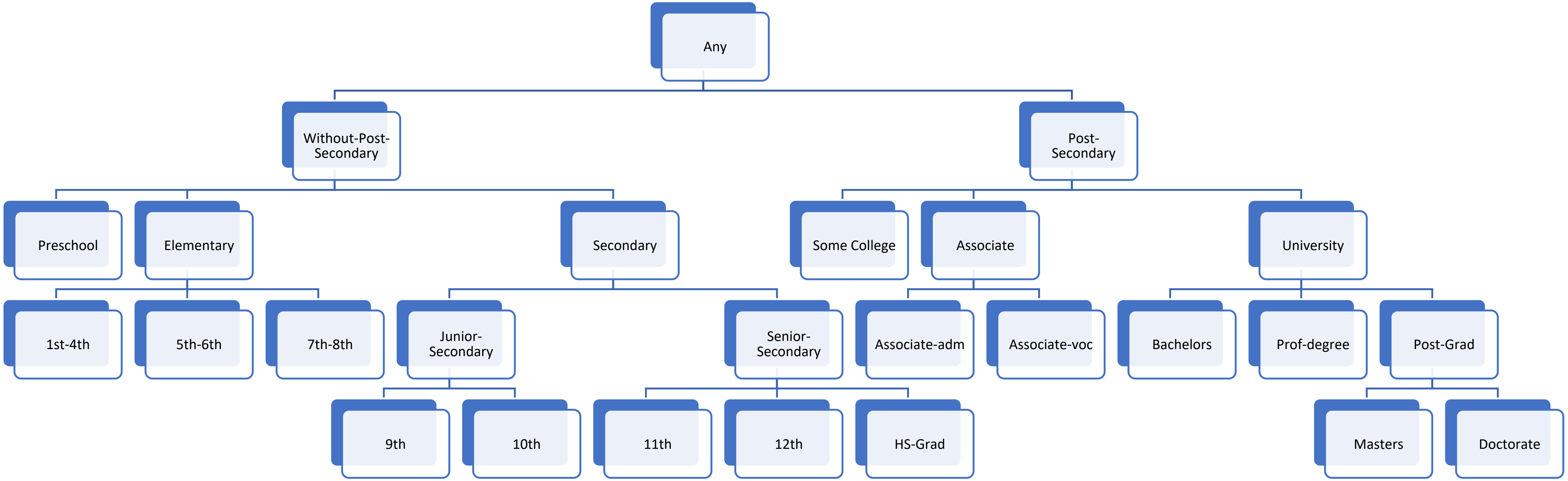


Pre-Processing

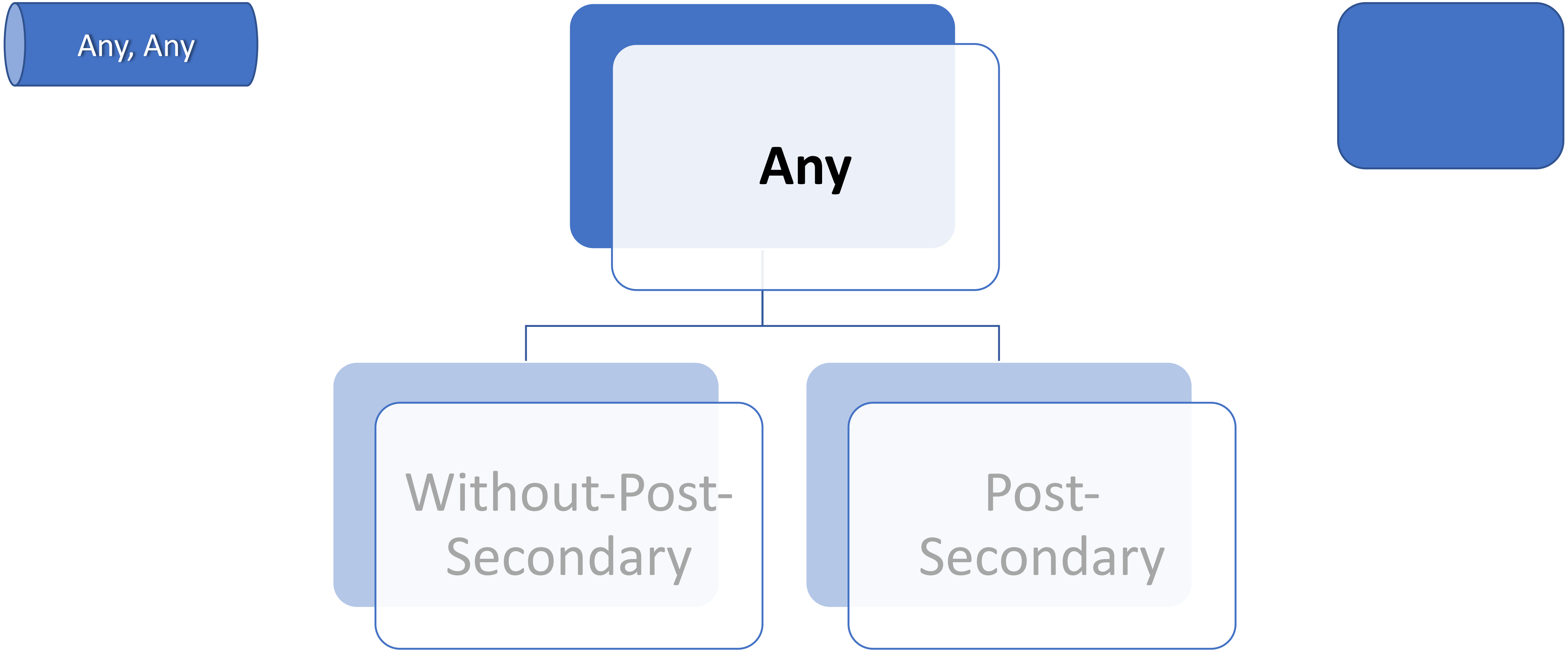
- Non QIDs are removed from the dataset
- QIDs and distinct values of SAs are grouped together and count calculated

Education	Gender	City	Income	Count
12th	Female	Orleans	<=50k	3
Bachelors	Female	Gloucester	>50k	4
Doctorate	Female	Gloucester	>50k	1
Bachelors	Female	Nepean	>50k	4
Associate	Male	Kanata	<=50k	4
11th	Male	Barrhaven	<=50k	2
Masters	Female	Perth	>50k	3

Pre-Processing – Building Path Maps



Pre-Processing – Building Path Maps



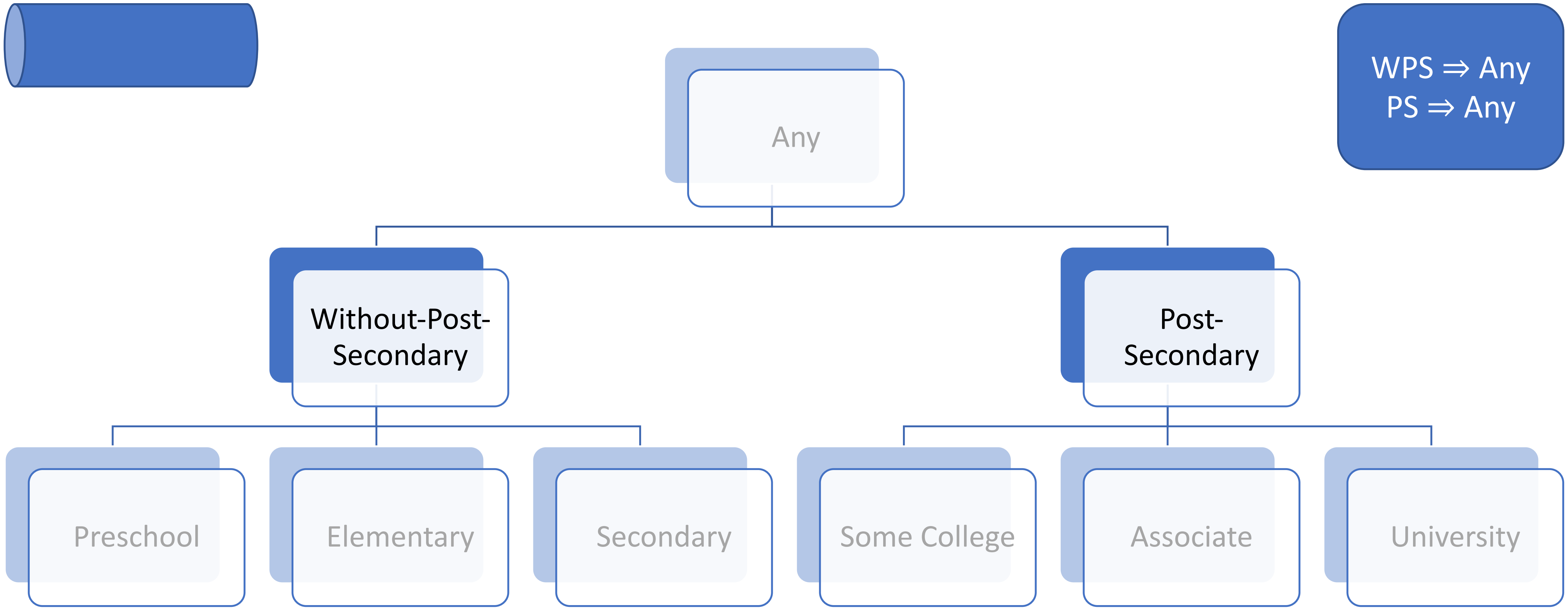
Any, Any

Any

Without-Post-Secondary

Post-Secondary

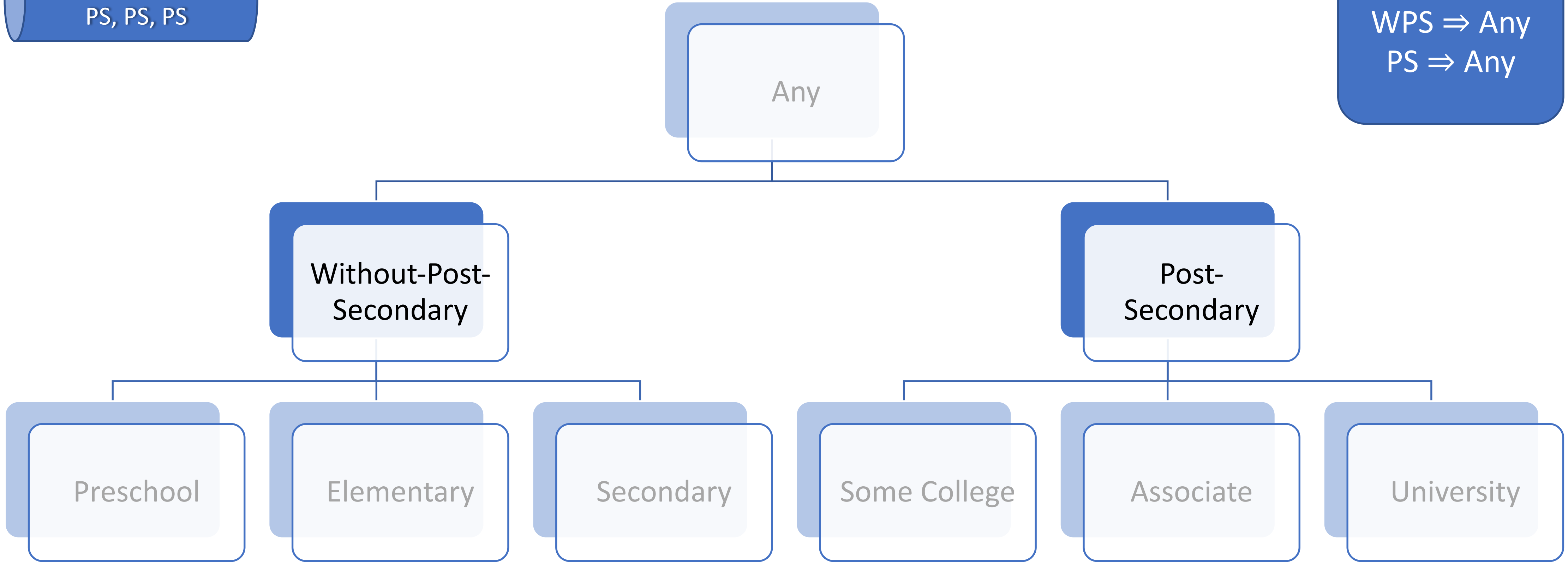
Pre-Processing – Building Path Maps



WPS ⇒ Any
PS ⇒ Any

Pre-Processing – Building Path Maps

WPS, WPS, WPS,
PS, PS, PS



Any

Without-Post-Secondary

Post-Secondary

Preschool

Elementary

Secondary

Some College

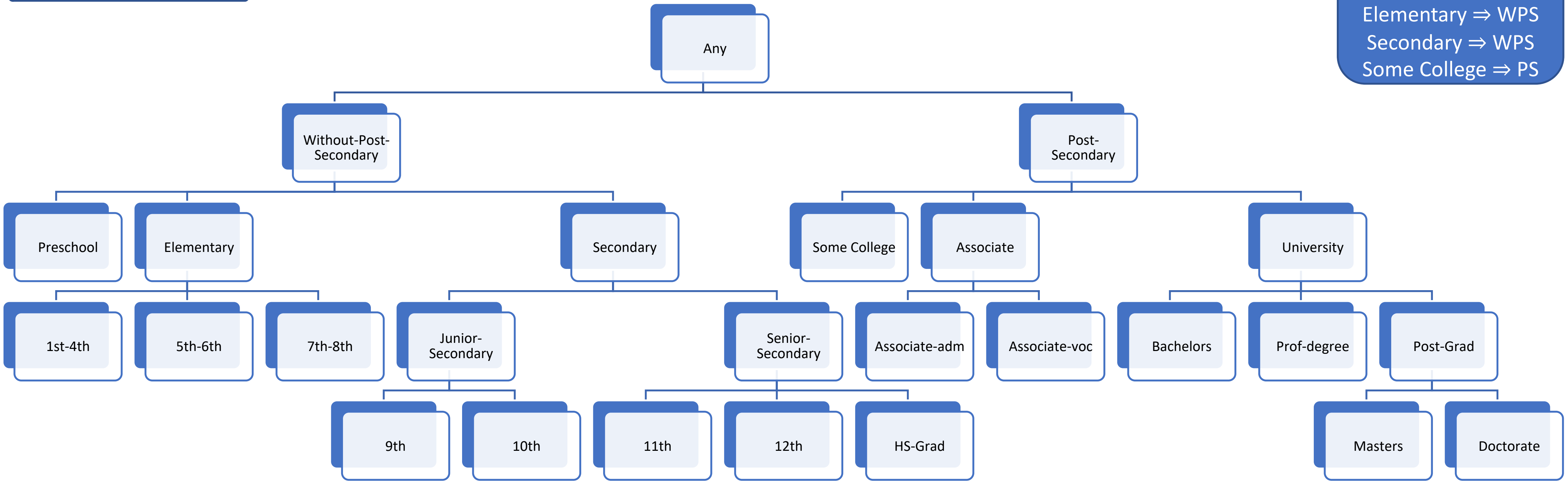
Associate

University

WPS ⇒ Any
PS ⇒ Any

Pre-Processing – Building Path Maps

WPS ⇒ Any
PS ⇒ Any
Preschool ⇒ WPS
Elementary ⇒ WPS
Secondary ⇒ WPS
Some College ⇒ PS



Any

Without-Post-Secondary

Post-Secondary

Preschool

Elementary

Secondary

Some College

Associate

University

1st-4th

5th-6th

7th-8th

Junior-Secondary

Senior-Secondary

Associate-adm

Associate-voc

Bachelors

Prof-degree

Post-Grad

9th

10th

11th

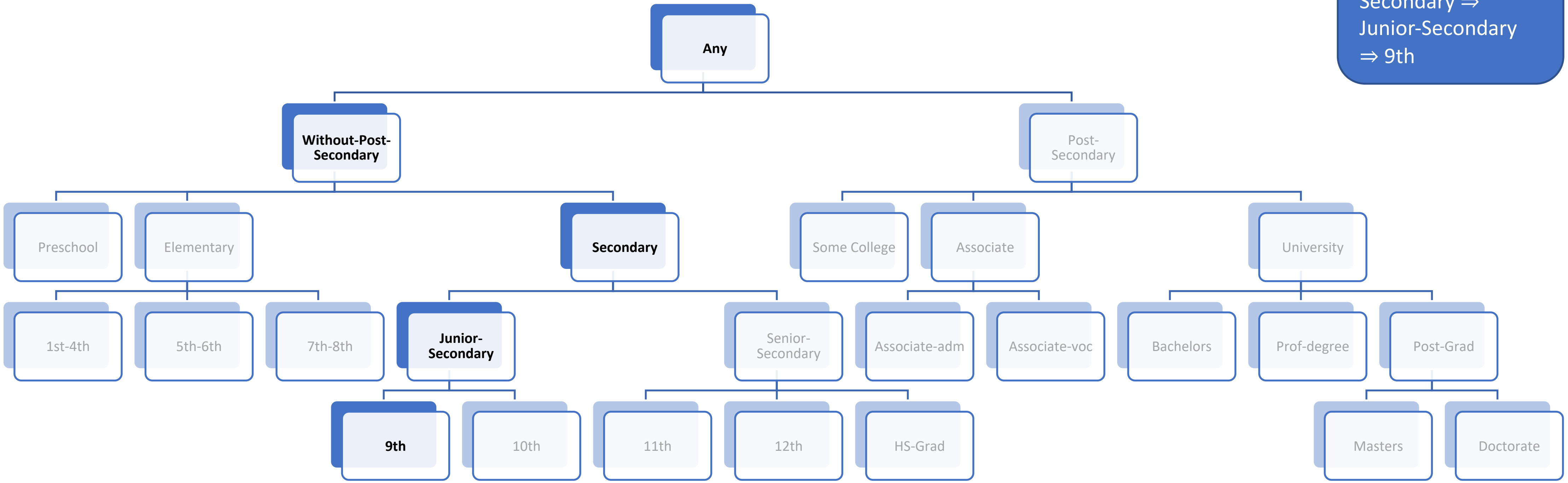
12th

HS-Grad

Masters

Doctorate

Pre-Processing – Building Path Maps



9th:
Any ⇒ WPS ⇒ Secondary ⇒ Junior-Secondary ⇒ 9th

Step 1 - Generalization

- Generalize all QIDs to the root of the anonymization level

Education	Gender	City	Income	Count	Aggregate
Any	Any	Any	<=50k	3	21
Any	Any	Any	>50k	4	
Any	Any	Any	>50k	1	
Any	Any	Any	>50k	4	
Any	Any	Any	<=50k	4	
Any	Any	Any	<=50k	2	
Any	Any	Any	>50k	3	

Step 1 – Pick Anonymization Level

- Generalize all QIDs to the root of the anonymization level
- For every anonymization level, calculate information gain and privacy loss

Education	Gender	City	Income	Count	Aggregate
WPS	Any	Any	<=50k	3	7
WPS	Any	Any	>50k	4	
PS	Any	Any	>50k	**1**	**14**
PS	Any	Any	>50k	**4**	
PS	Any	Any	<=50k	**4**	
PS	Any	Any	<=50k	**2**	
PS	Any	Any	>50k	**3**	

Step 2 – Pick Anonymization Level

- Generalize all QIDs to the root of the anonymization level
- For every anonymization level, calculate information gain and privacy loss

Education	Gender	City	Income	Count	Aggregate
Any	Any	East	<=50k	3	8
Any	Any	East	>50k	4	
Any	Any	East	>50k	1	
Any	Any	**West**	>50k	**4**	**13**
Any	Any	**West**	<=50k	**4**	
Any	Any	**West**	<=50k	**2**	
Any	Any	**West**	>50k	**3**	

Step 3 – Score Best Option

- Anonymization level values are aggregated for every partition
- Aggregations are merged into one-row table with the totals

Education	Gender	City	Income	Count	Edu_Agg	City_Agg
WPS	Any	East	<=50k	3	7	8
WPS	Any	East	>50k	4		
PS	Any	East	>50k	1	14	
PS	Any	West	>50k	4		13
PS	Any	West	<=50k	4		
PS	Any	West	<=50k	2		
PS	Any	West	>50k	3		

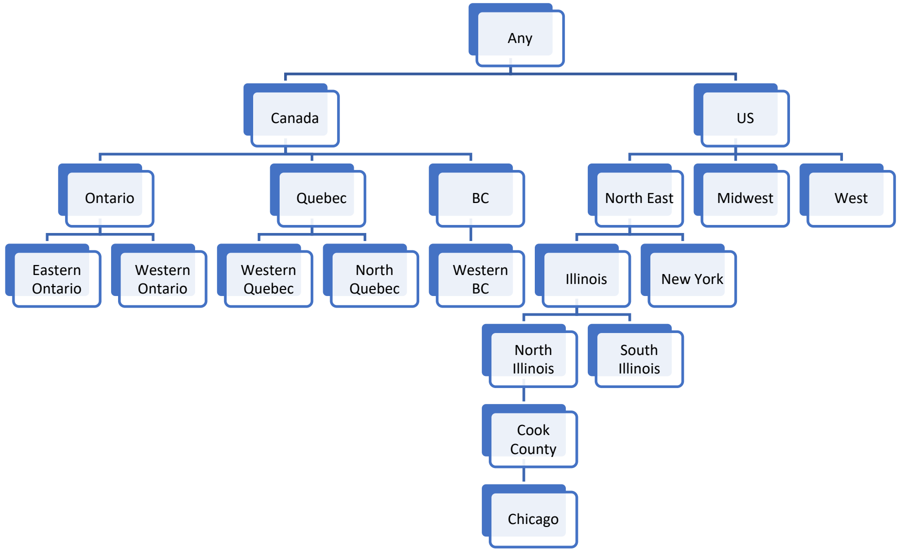
Best Score: City

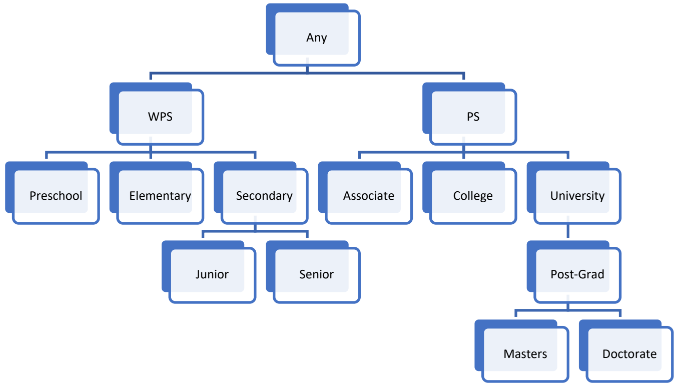
Orleans: Any ⇒ Canada ⇒ Ontario ⇒ Eastern Ontario ⇒ Greater Ottawa Area ⇒ Ottawa East ⇒ Orleans

Chicago: Any ⇒ United States ⇒ North East ⇒ Illinois ⇒ North Illinois ⇒ Cook County ⇒ Chicago

Step 3 – Score Best Option

- Anonymization level values are aggregated for every partition
- Aggregations are merged into one-row table with the totals
- Calculate score and re-iterate





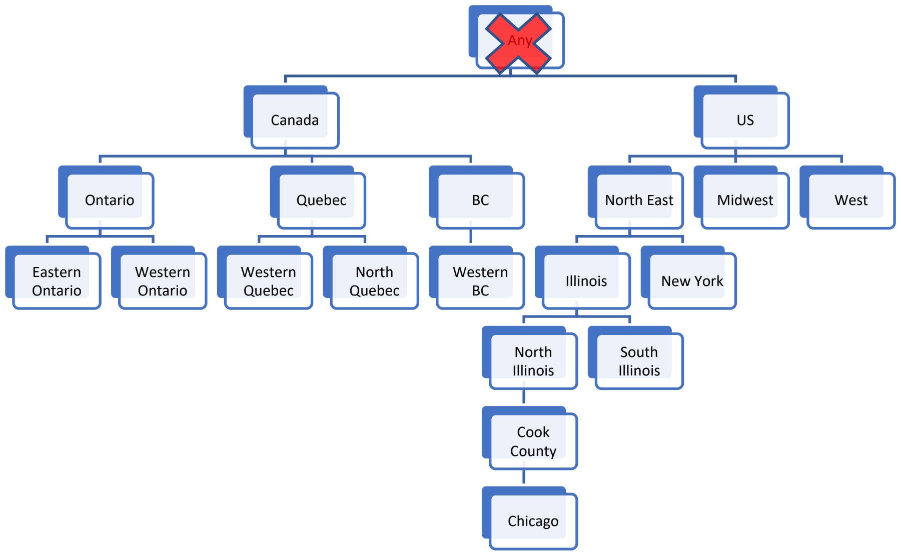
Best Score: City

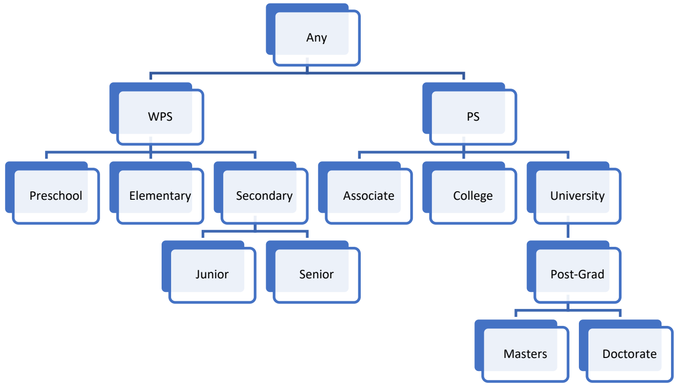
Orleans: ~~**Any**~~ ⇒ Canada ⇒ Ontario ⇒ Eastern Ontario ⇒ Greater Ottawa Area ⇒ Ottawa East ⇒ Orleans

Chicago: ~~**Any**~~ ⇒ United States ⇒ North East ⇒ Illinois ⇒ North Illinois ⇒ Cook County ⇒ Chicago

Step 3 – Score Best Option

- Anonymization level values are aggregated for every partition
- Aggregations are merged into one-row table with the totals
- Calculate score and re-iterate





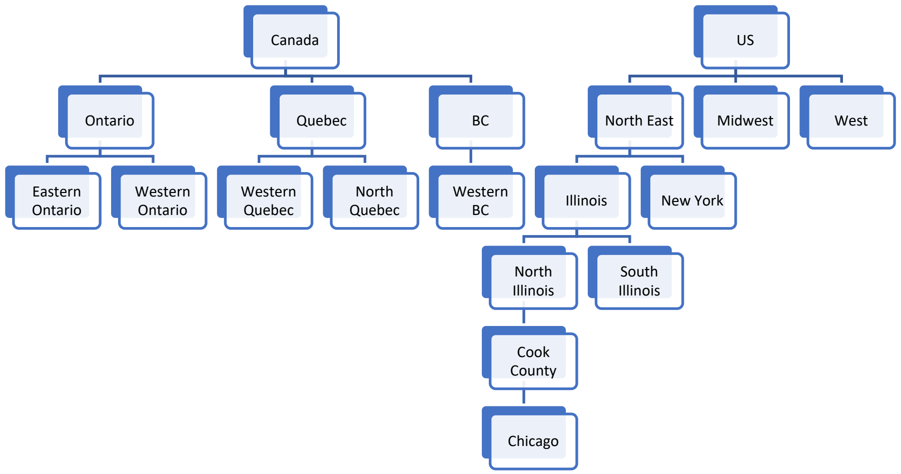
Best Score: City

Orleans: ~~**Any**~~ ⇒ Canada ⇒ Ontario ⇒ Eastern Ontario ⇒ Greater Ottawa Area ⇒ Ottawa East ⇒ Orleans

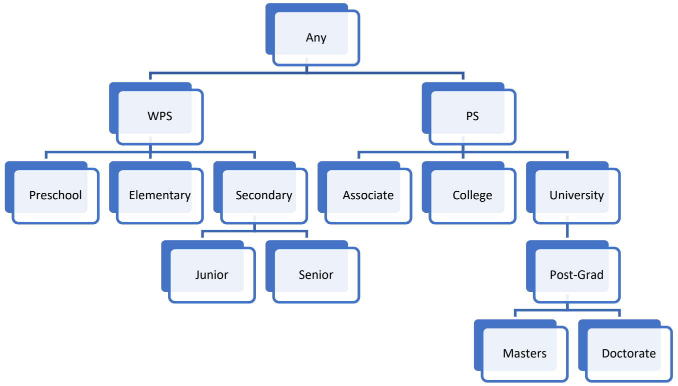
Chicago: ~~**Any**~~ ⇒ United States ⇒ North East ⇒ Illinois ⇒ North Illinois ⇒ Cook County ⇒ Chicago

Step 3 – Score Best Option

- Anonymization level values are aggregated for every partition
- Aggregations are merged into one-row table with the totals
- Calculate score and re-iterate



Canada
Ontario
Eastern Ontario
Western Ontario
Quebec
Western Quebec
North Quebec
BC
Western BC
US
North East
Illinois
North Illinois
Cook County
Chicago
South Illinois
New York
Midwest
West



Any
WPS
Preschool
Elementary
Secondary
Junior
Senior
PS
Associate
College
University
Post-Grad
Masters
Doctorate

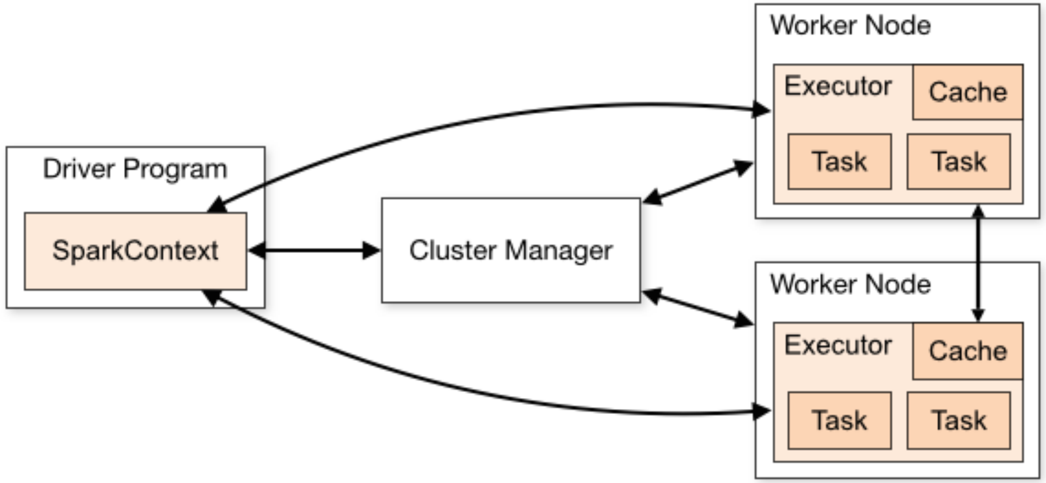
Best Score: City

Orleans: ~~**Any** ⇒~~ Canada ⇒ Ontario ⇒ Eastern Ontario ⇒ Greater Ottawa Area ⇒ Ottawa East ⇒ Orleans

Chicago: ~~**Any** ⇒~~ United States ⇒ North East ⇒ Illinois ⇒ North Illinois ⇒ Cook County ⇒ Chicago

Enhancing performance

- Apache Spark is a fast and general-purpose cluster computing system
- Maximum partitions set to p where p is number of processors
- Prefer tail recursion over looping for code that runs on Spark
- Minimize aggregations to maximum 1 per iteration
- Partition over ID

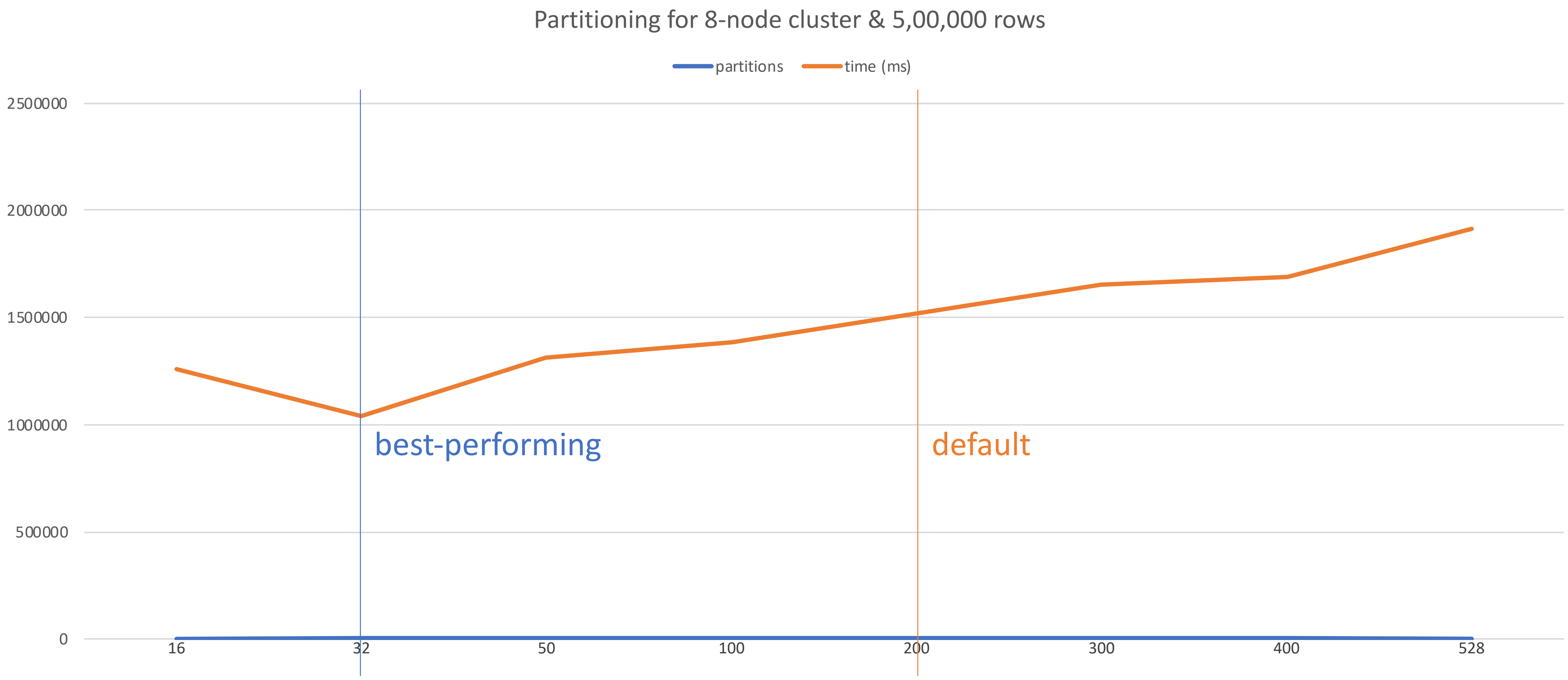


Test Environment

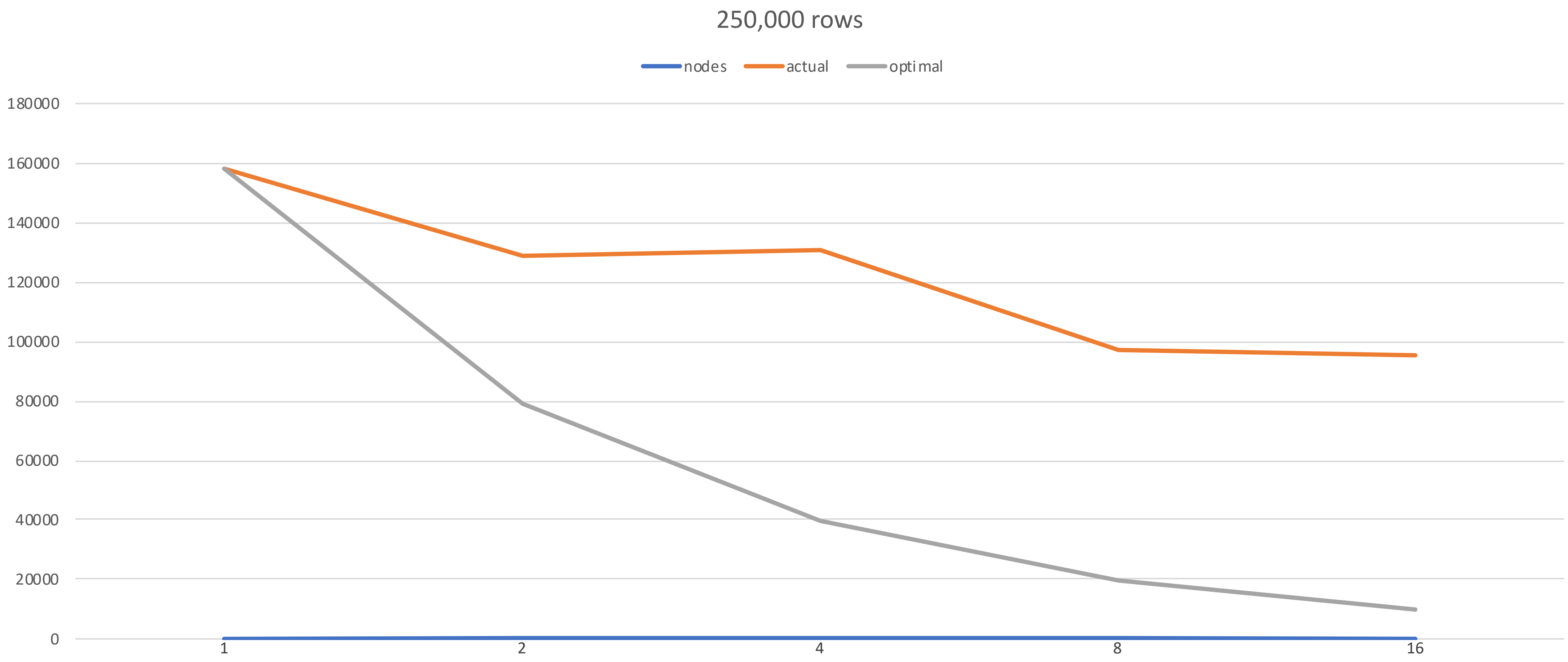
- OpenStack configuration with 32 GB disk space, 8 GB RAM, 4 vCPU per node
- Ran tests for k=100 over 1, 2, 4, 8 and 16 nodes
- Dataset sizes: 250,000 rows, 5 million rows and 10 million rows
- Spark and Java installed on every node
- Public/private keys added to every node and hosts file updated

Number of Partitions

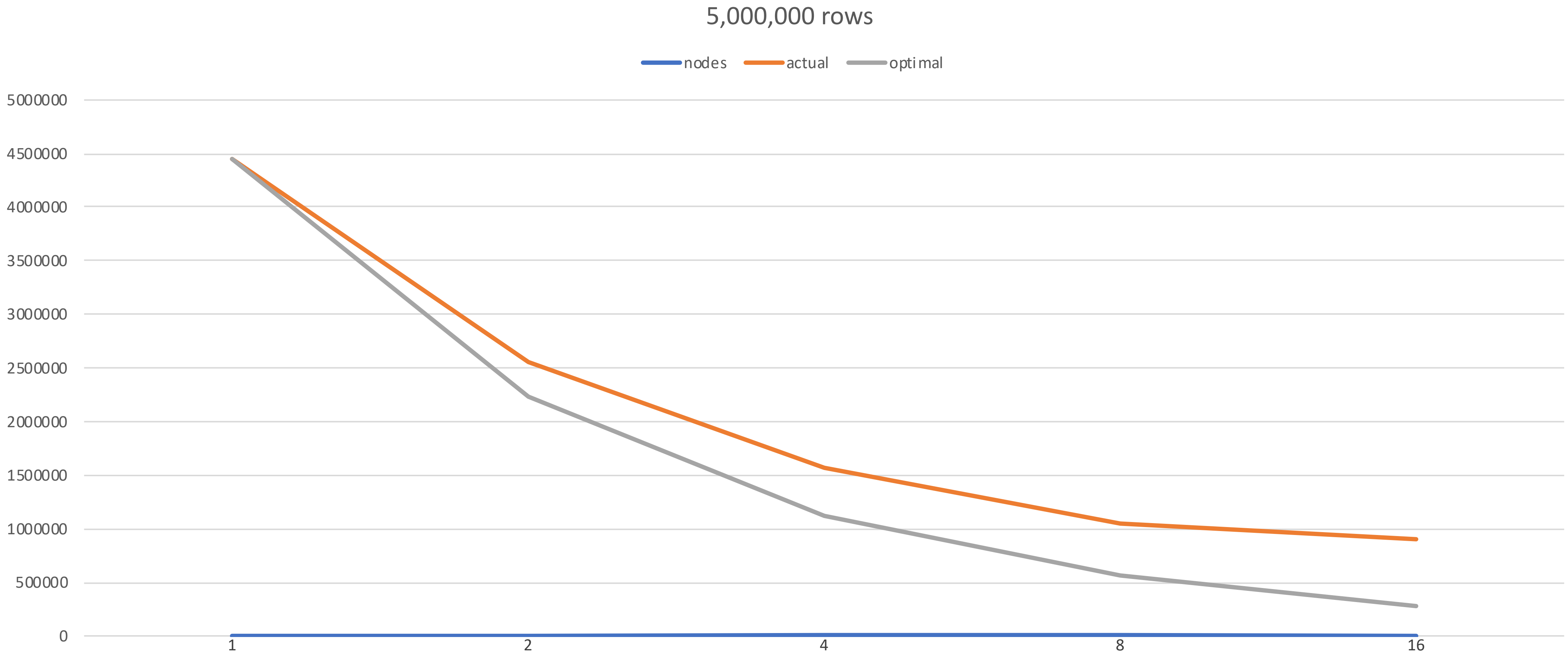
Set number of partitions to number of worker cores



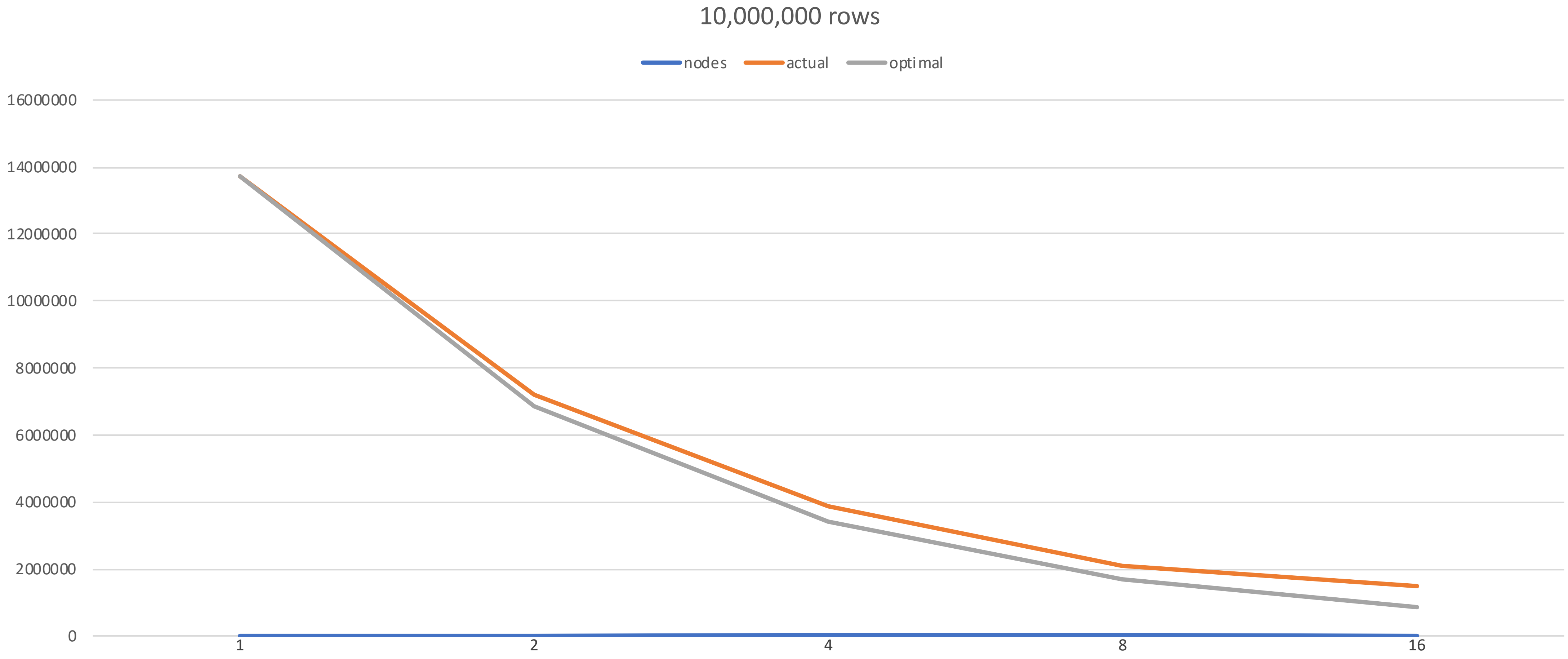
Test Results



Test Results

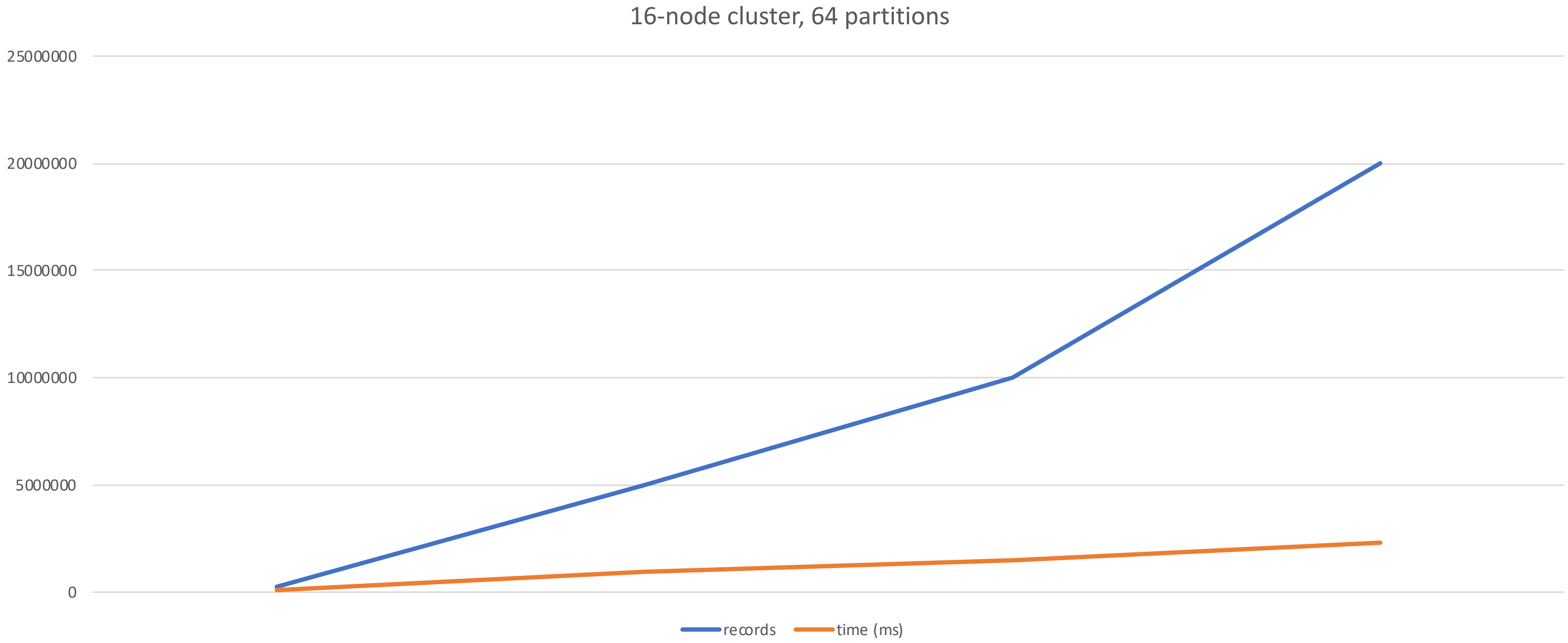


Test Results



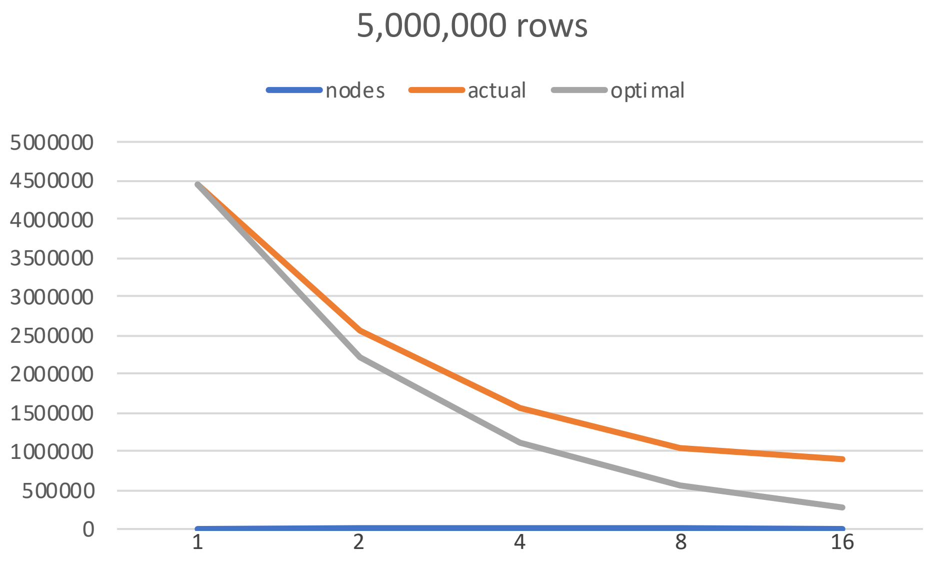
Performance by dataset size

100% increase in dataset size only resulted in 55-65% increase in time

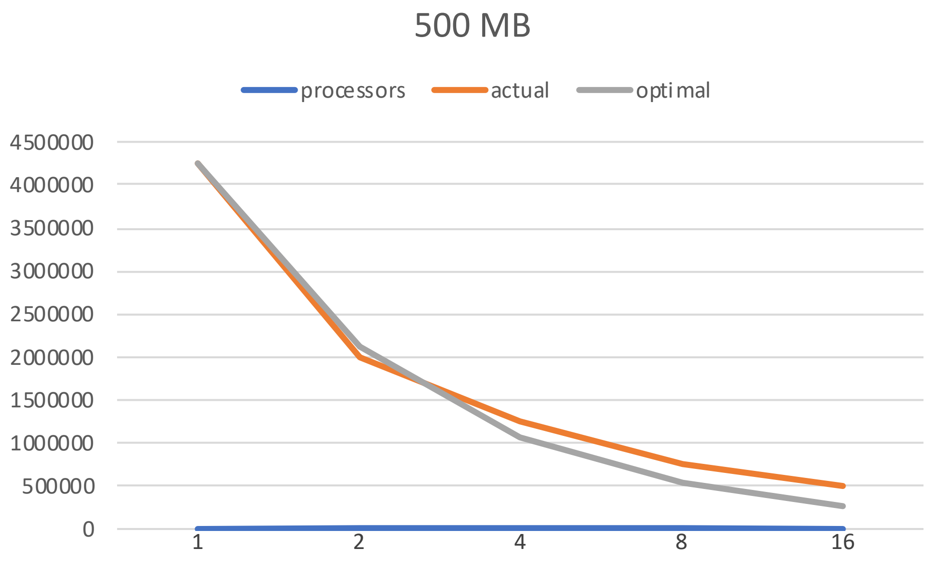


Comparison with Original Paper

My Implementation

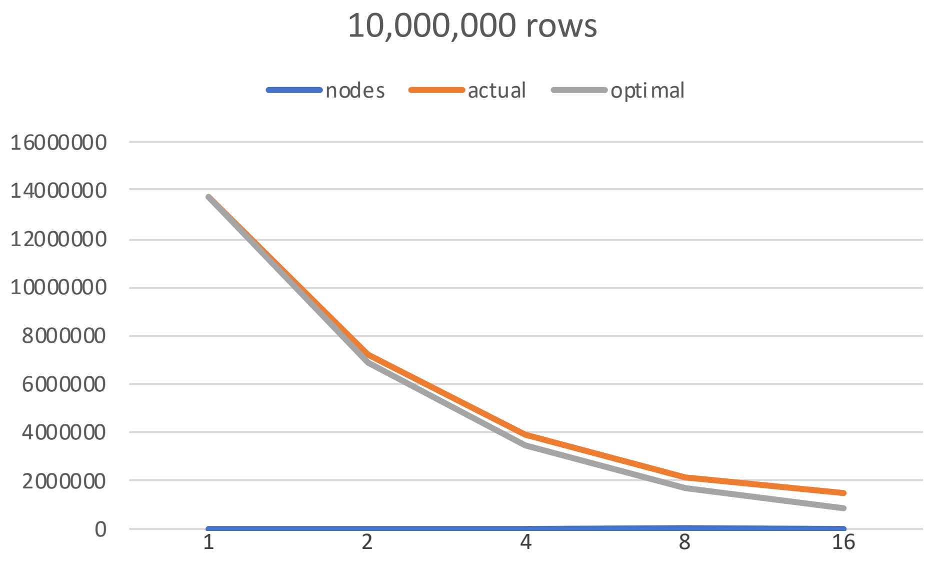


Original Paper

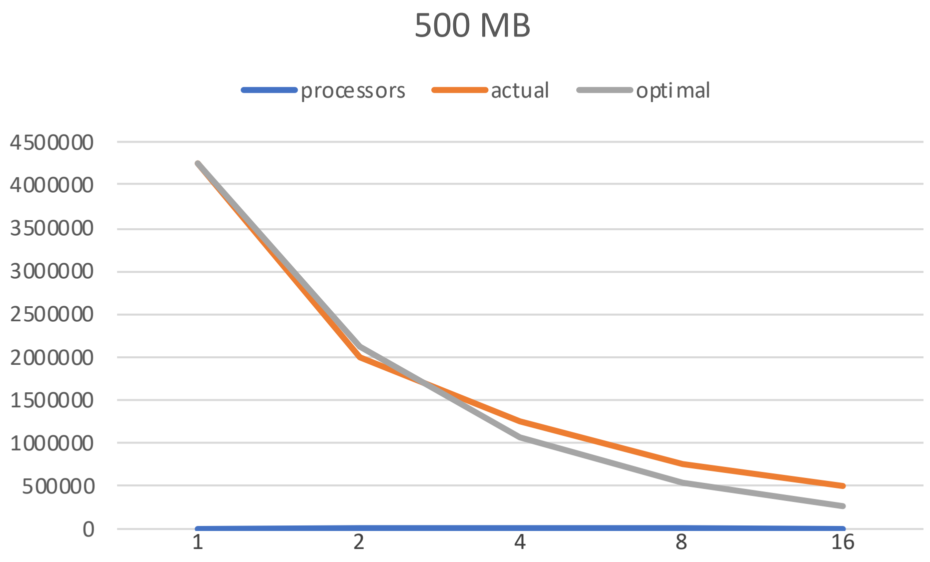


Comparison with Original Paper

My Implementation



Original Paper



Questions

- What’s the difference between Quasi-Identifiers and Sensitive Attributes?
- What change contributed the most to performance improvement?
- What should be the number of partitions compared to number of processors?