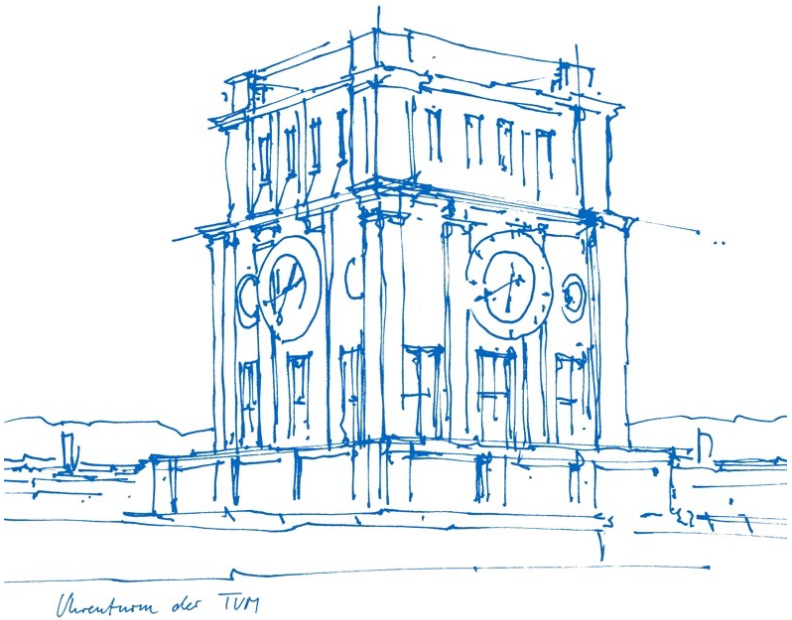
Chair of Database Systems
Departement of Informatics
Technical University of Munich

Column Imprints: A Secondary Index Structure

Seminar: Techniques for implementing main memory database systems



Aikaterini Intzevidou (MSc)
katerina.intzevidou@tum.de

December 10th, Munich

What are column imprints?

- Secondary index structure
  - Columnar databases
  - Beside the primary index
  - Bitmaps, Zone maps, etc.
- Space efficient
- Exploit local clustering

Building column imprints

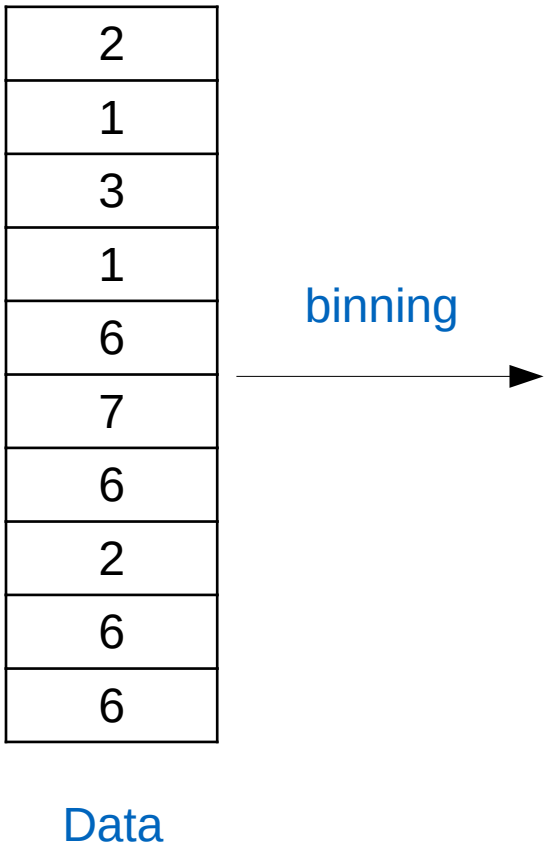
- Sample values from the column
- Create (equi-width or equi-depth) histogram → bin ranges
- Map cacheline of data to bins
  - Bit vectors with #bits equal to #bins
  - Set bit to 1 if value falls into the corresponding bin
- Boom! Index ready

Building column imprints

2
1
3
1
6
7
6
2
6
6

Data

Building column imprints

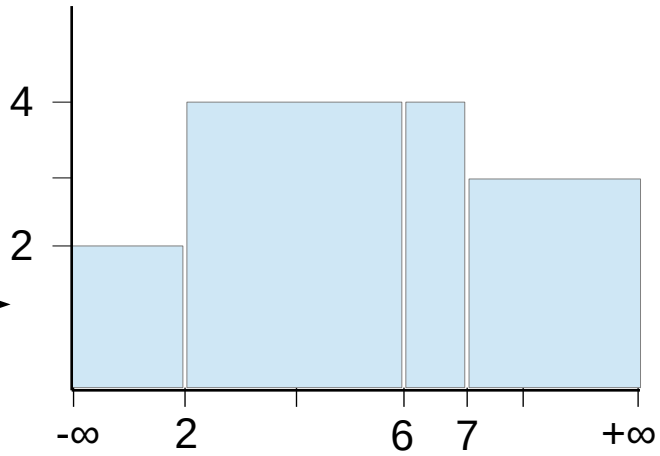


Building column imprints

2
1
3
1
6
7
6
2
6
6

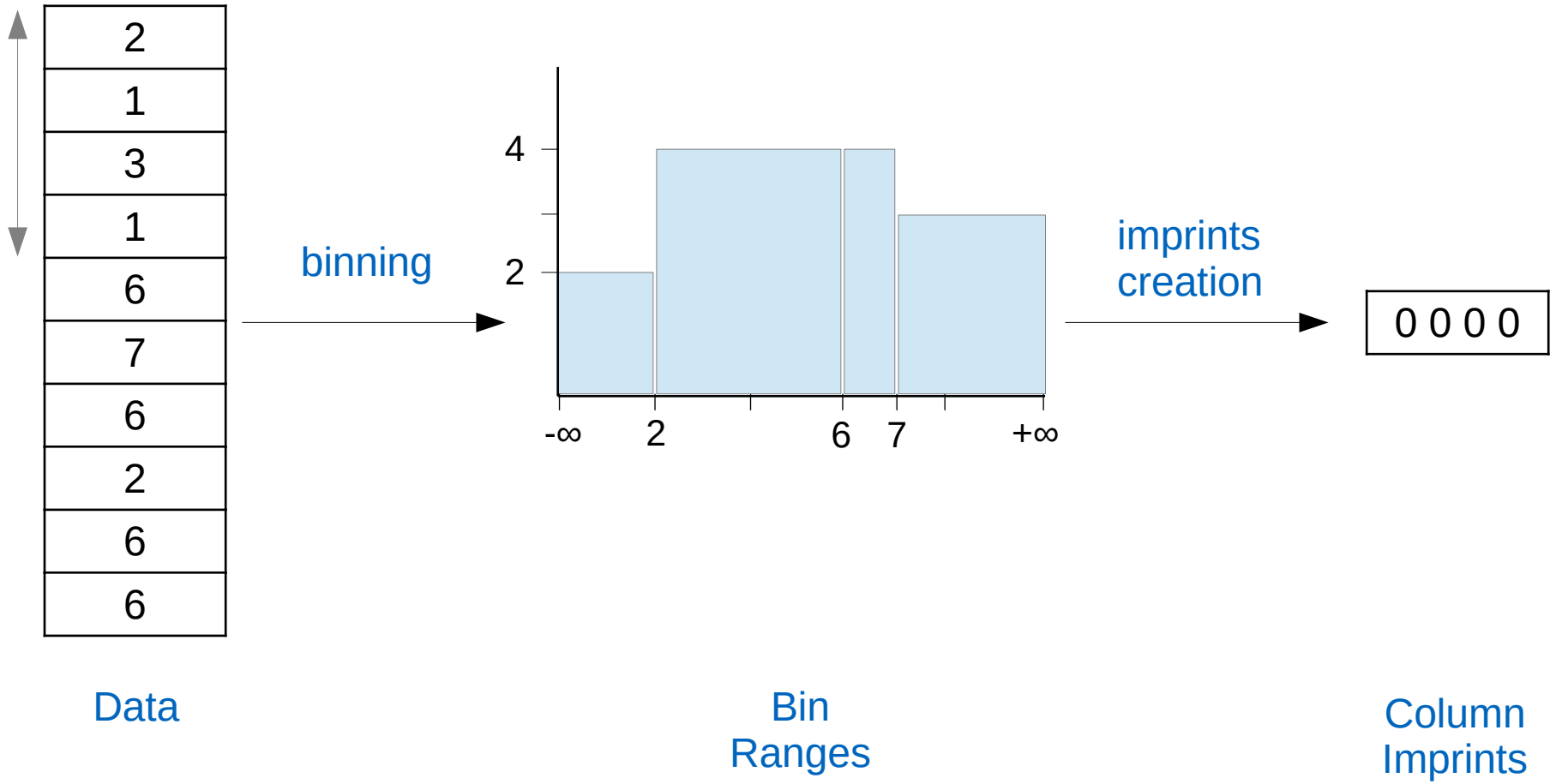
Data

binning

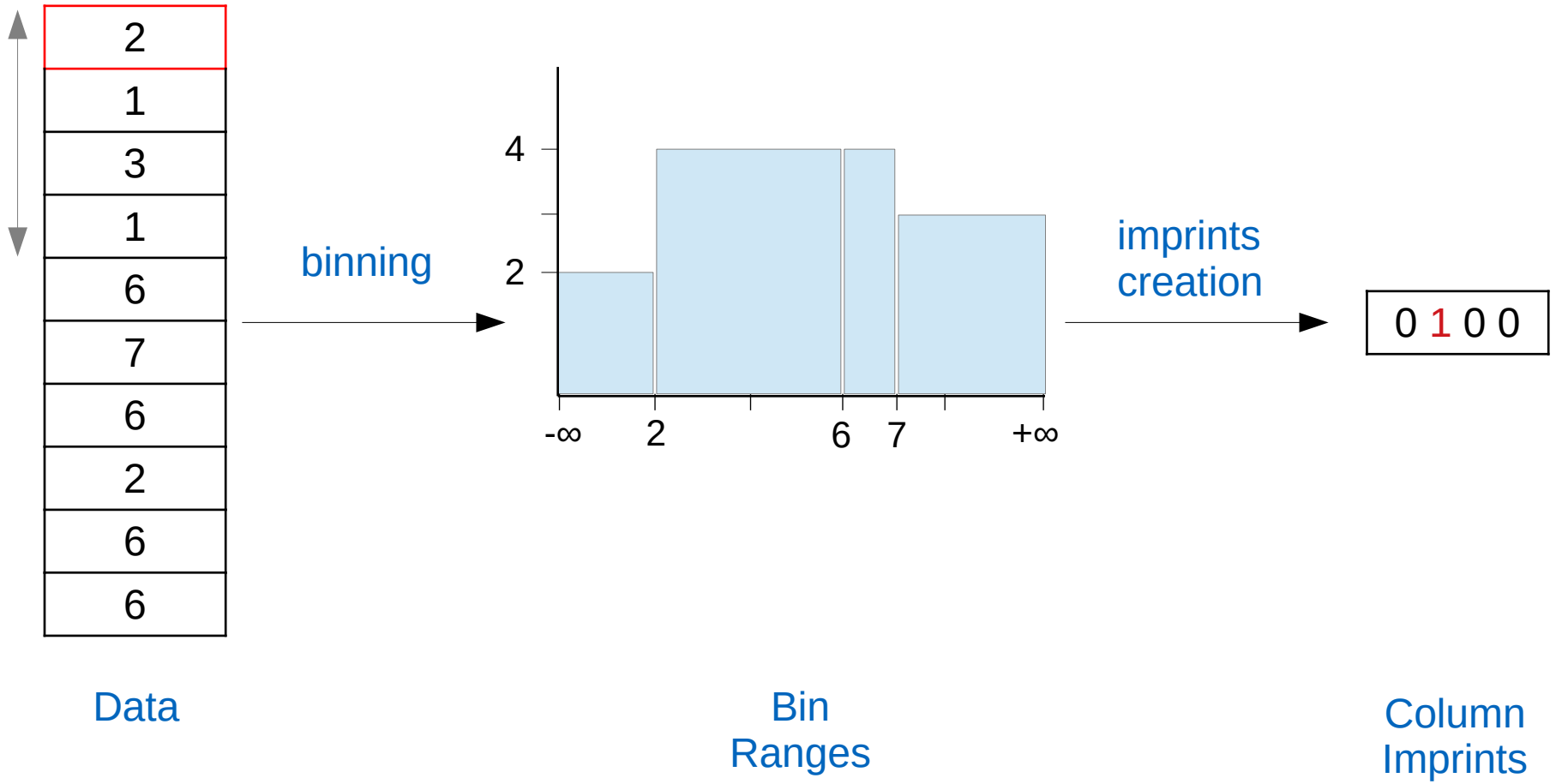


Bin Ranges

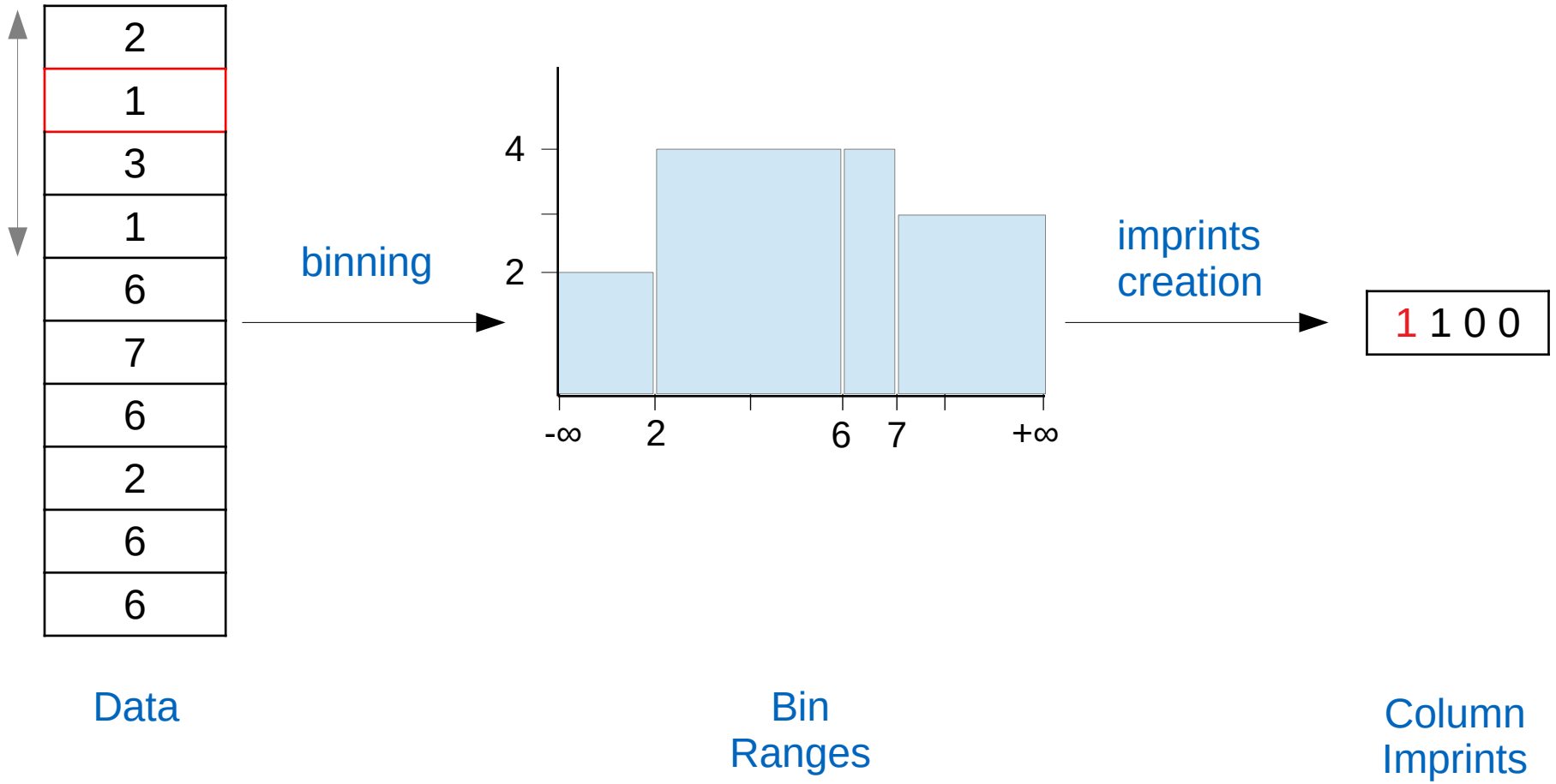
Building column imprints



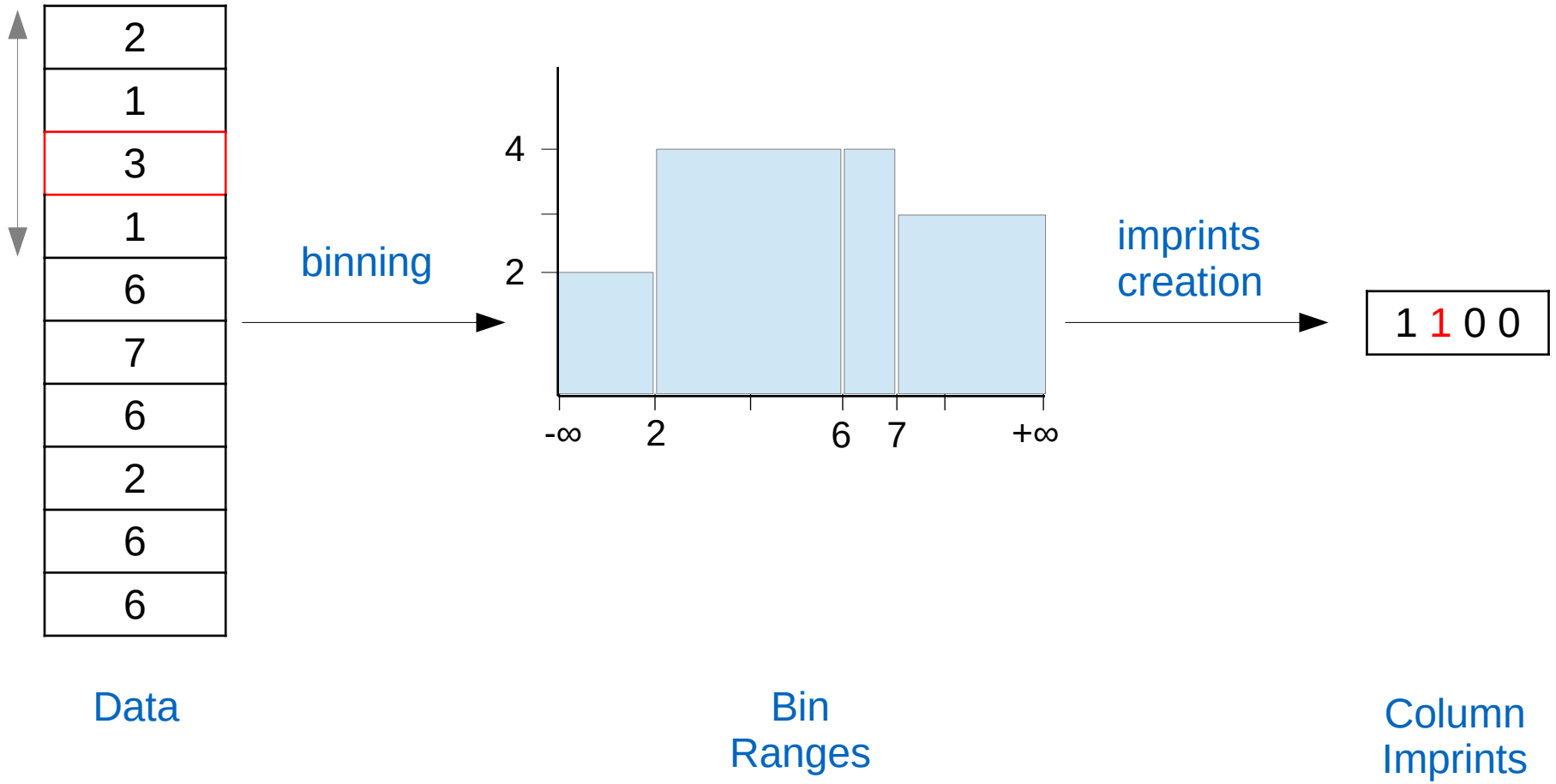
Building column imprints



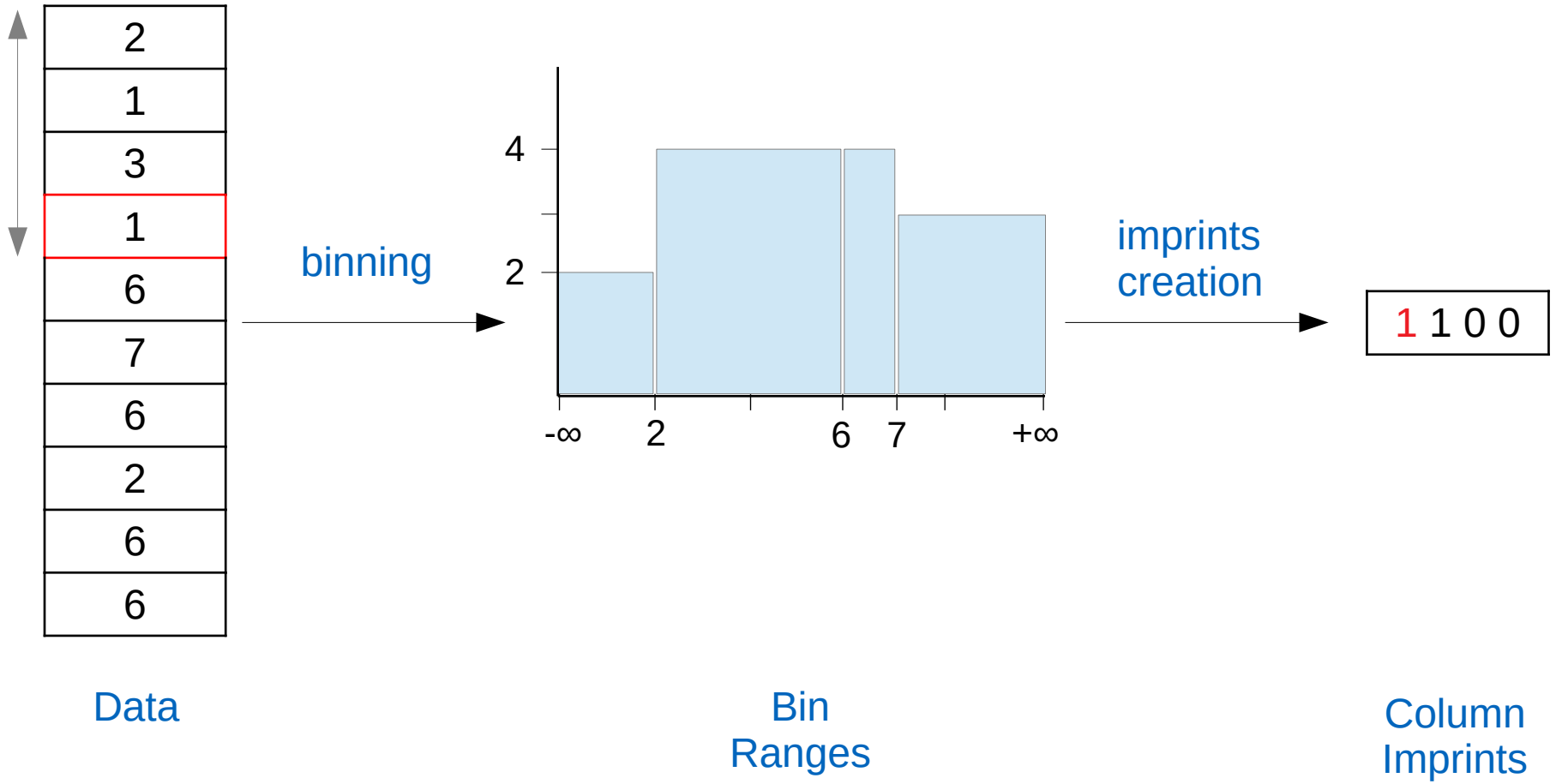
Building column imprints



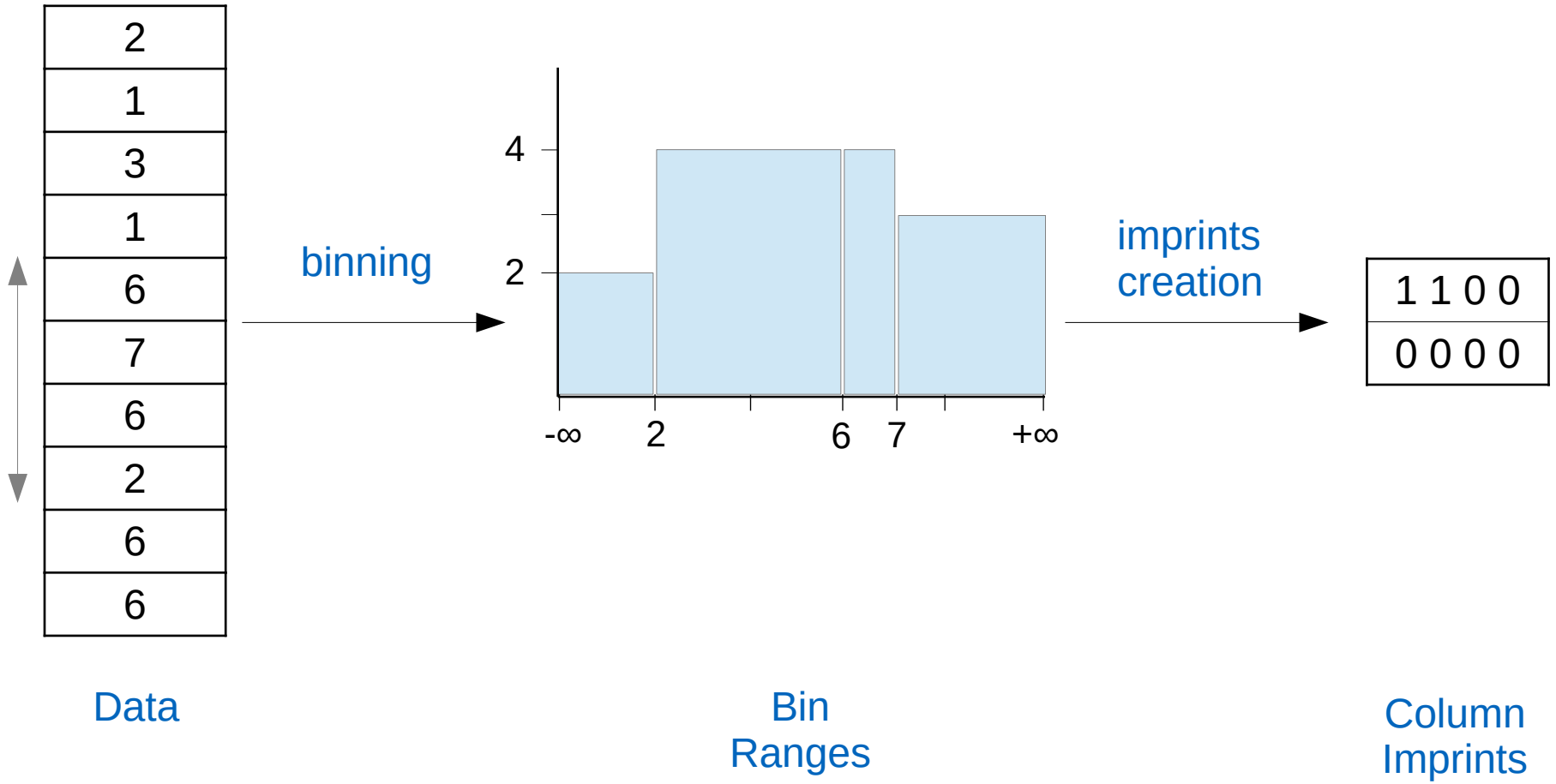
Building column imprints



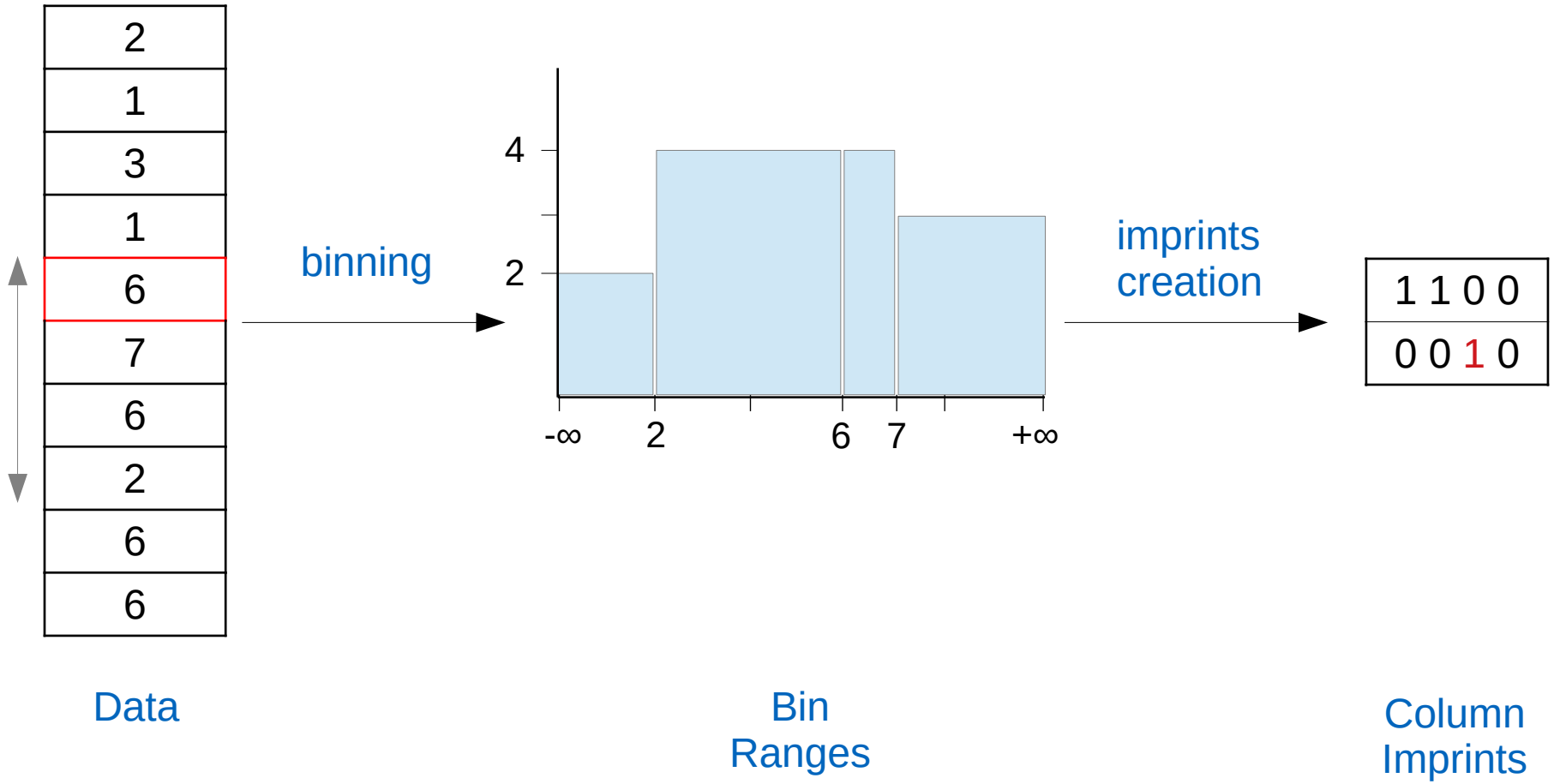
Building column imprints



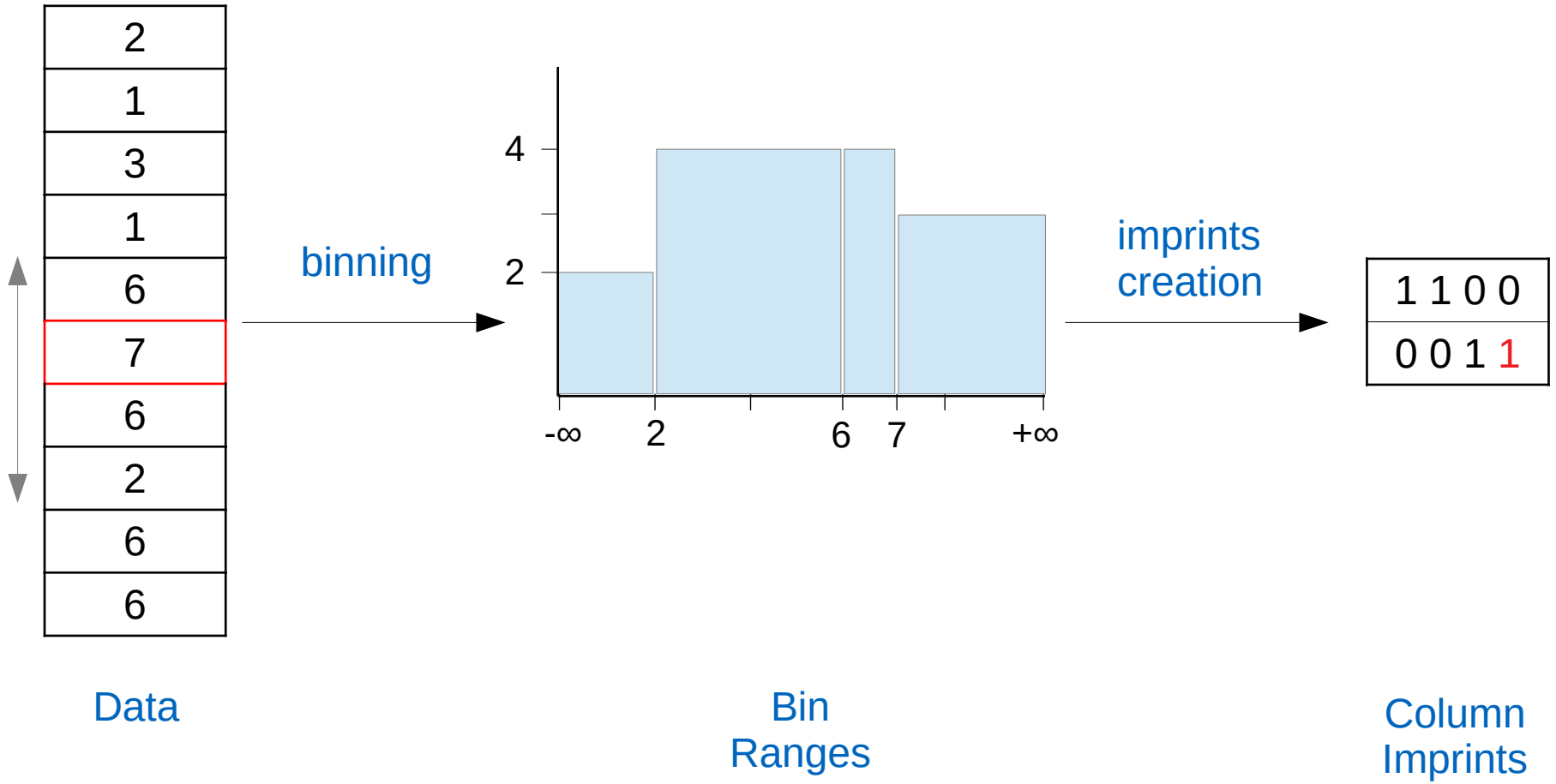
Building column imprints



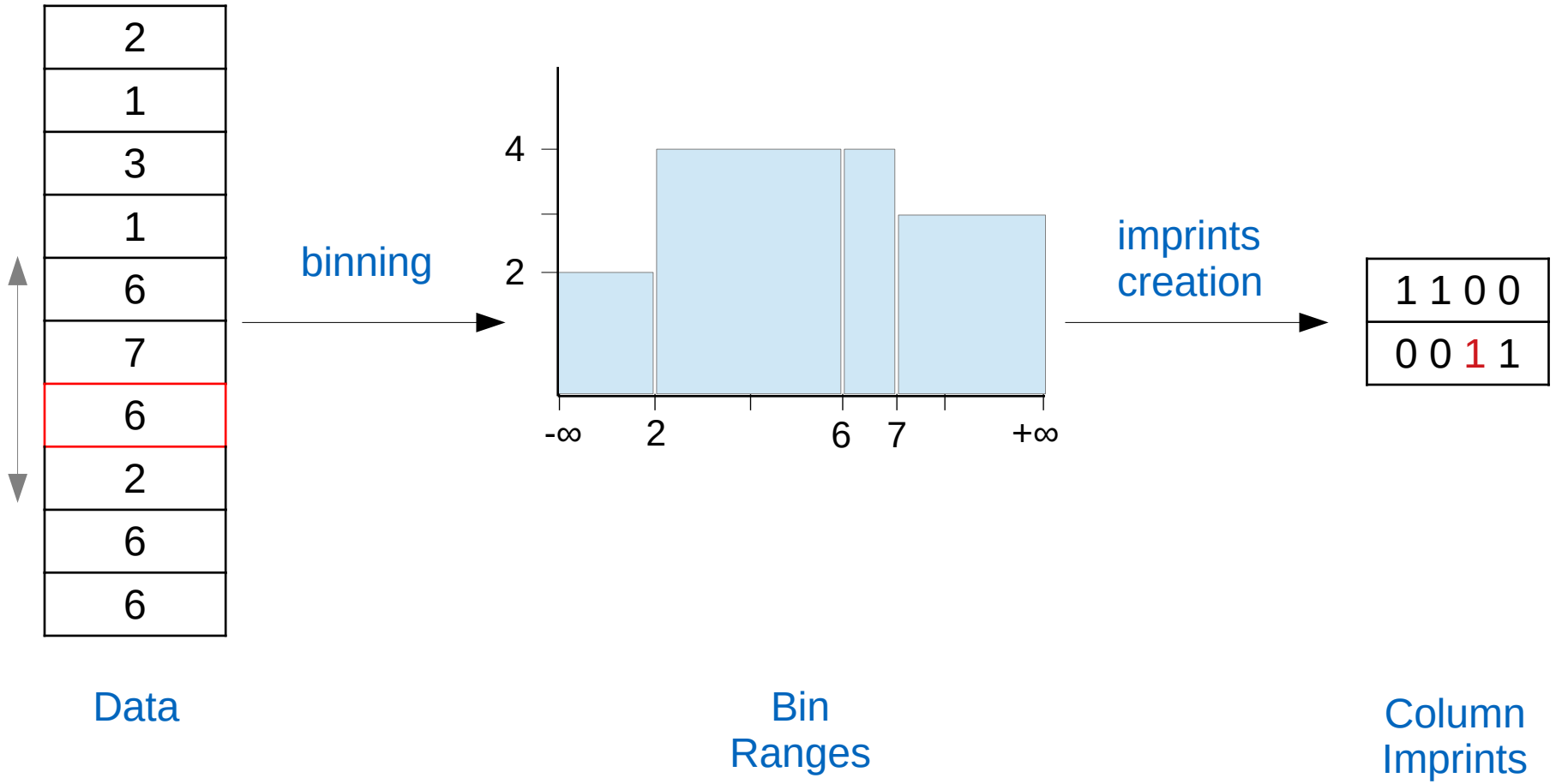
Building column imprints



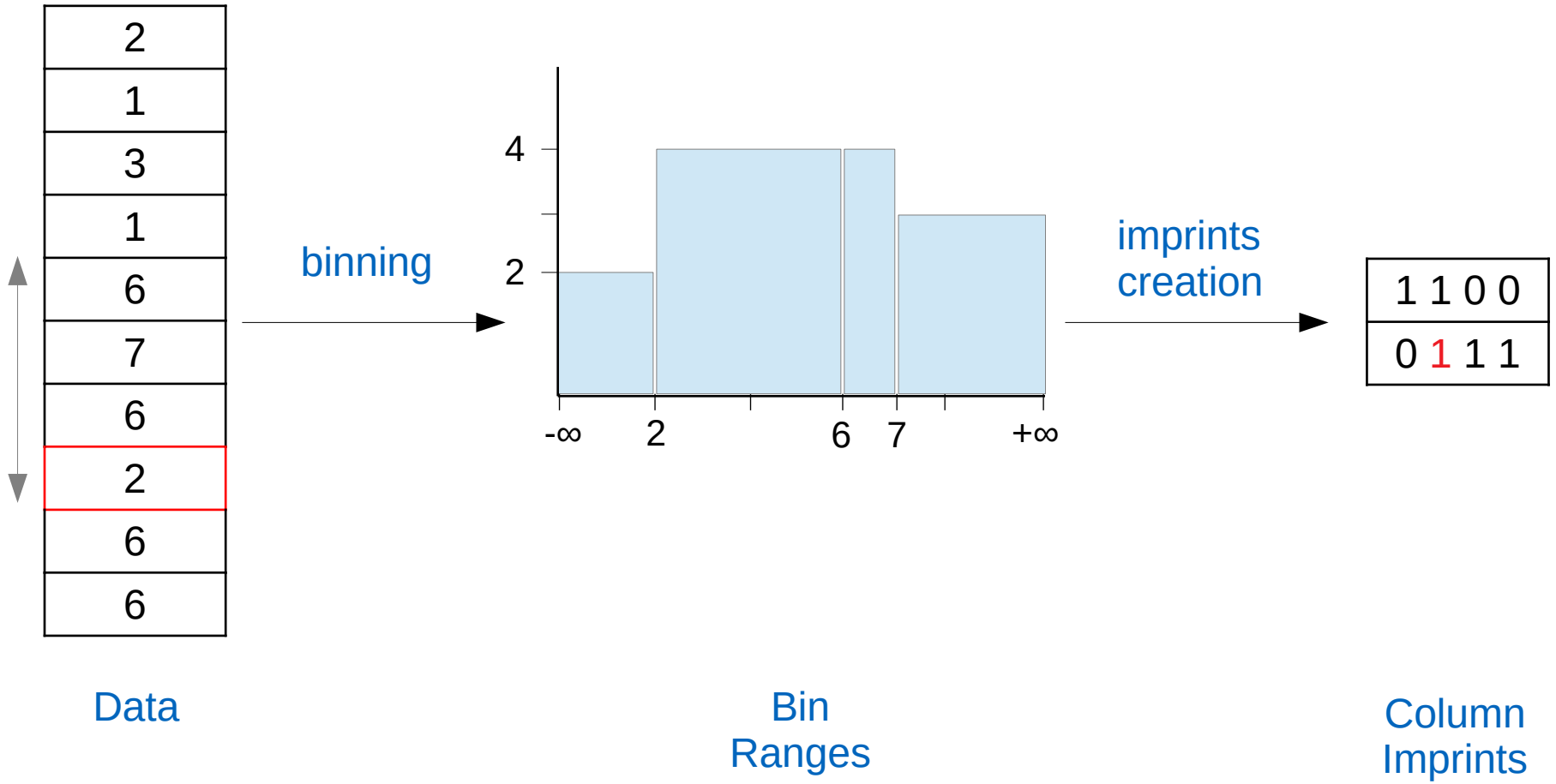
Building column imprints



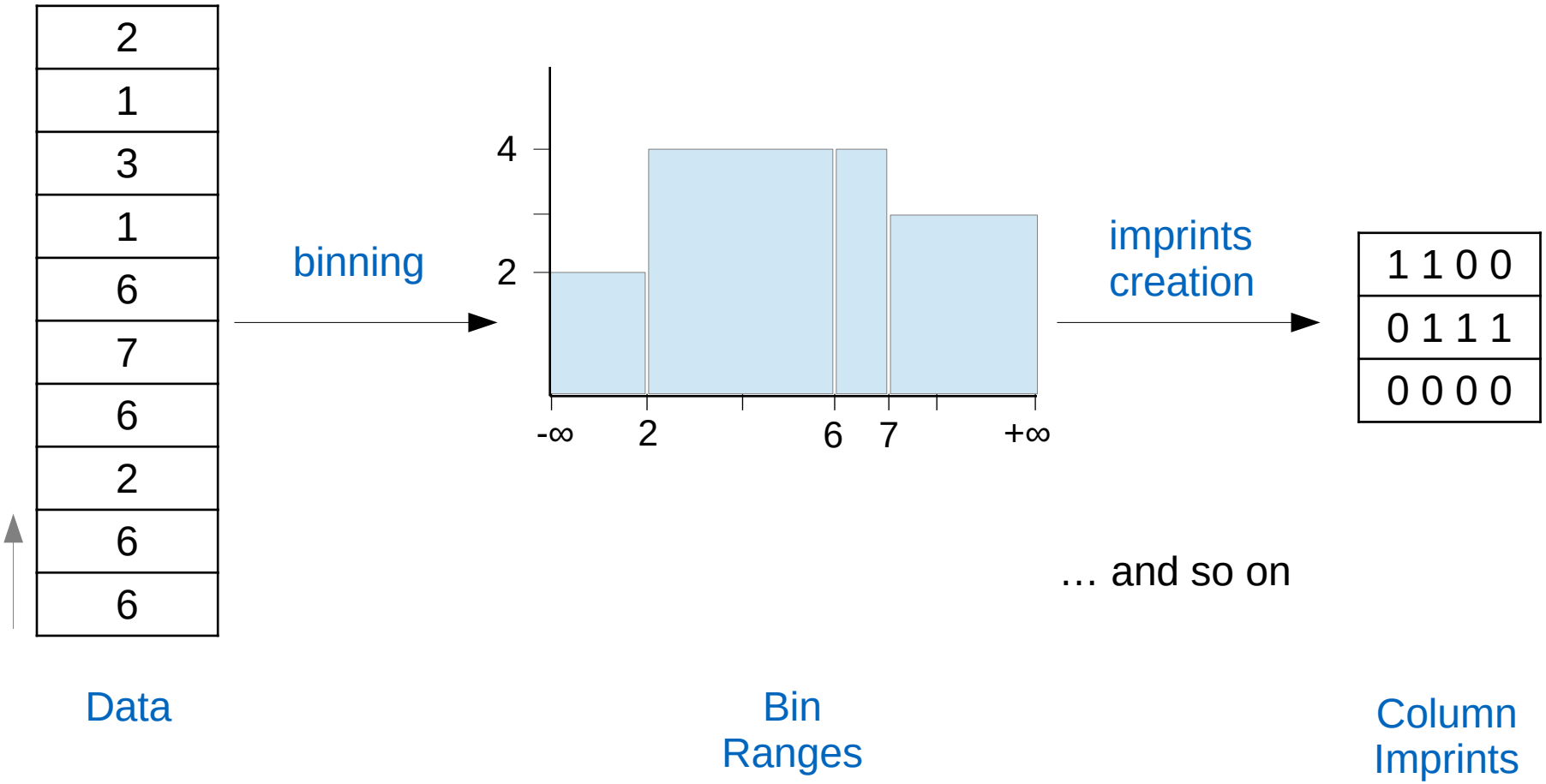
Building column imprints



Building column imprints



Building column imprints



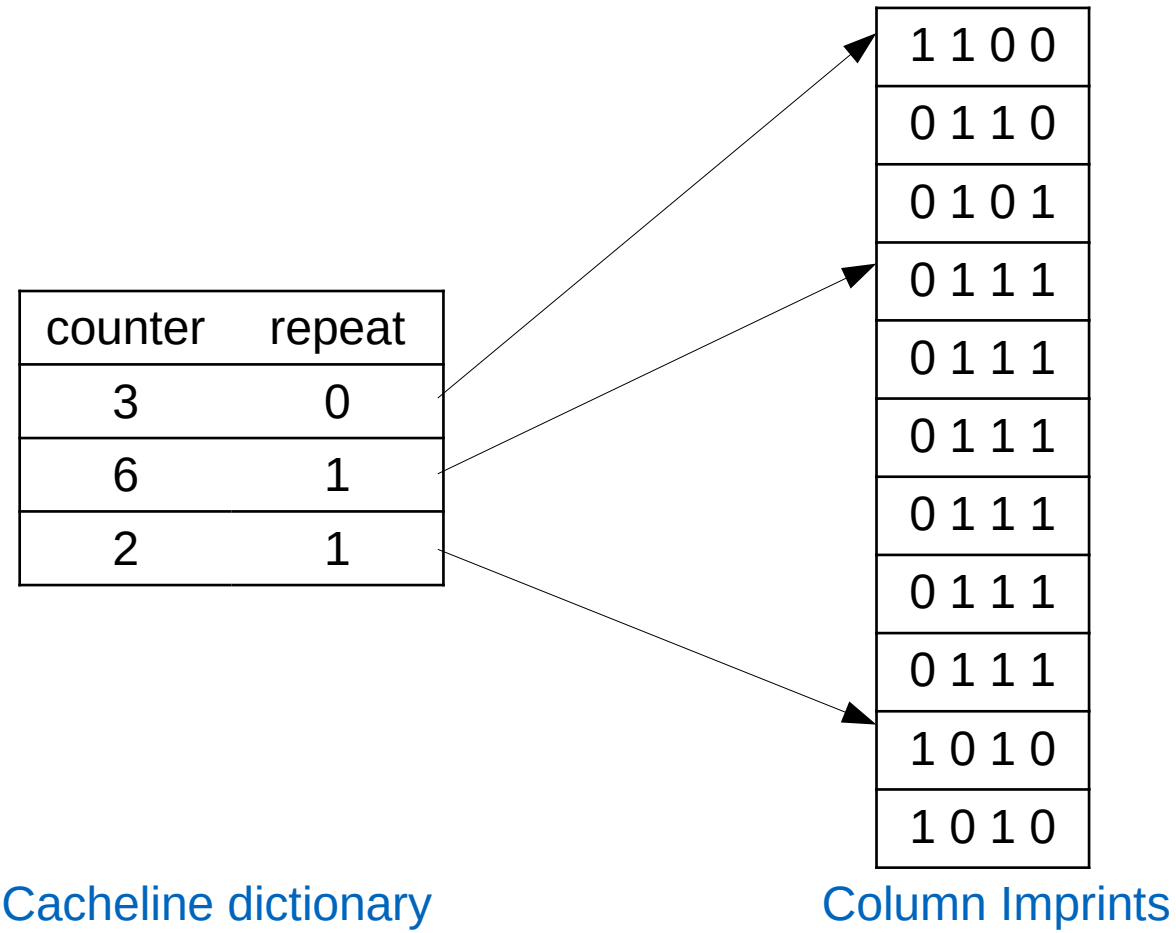
Building column imprints

Our column imprints are now created!

Can we make them better with compression?

Compression

The **cacheline dictionary**, keeps a counter and a repeat flag



Updating column imprints

- Bit vector is created for each new value
- Bits are set to 1, getting the corresponding bins, as in the building process
- New imprints are appended in the column imprints structure

Read intensive databases → Rarely happen + Done in batches

Querying

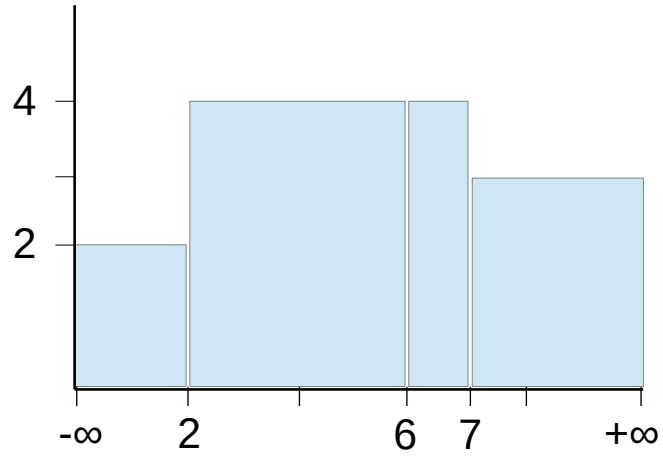
- Create bit vector for the query
- Compare to imprint vectors
- Return cachelines with matching values

Querying

Range query: [1, 3]

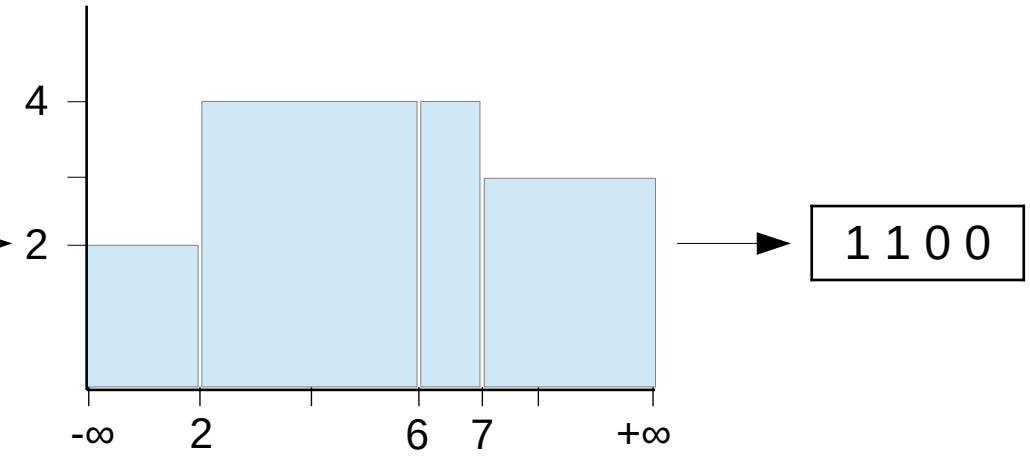
Querying

Range query: [1, 3]



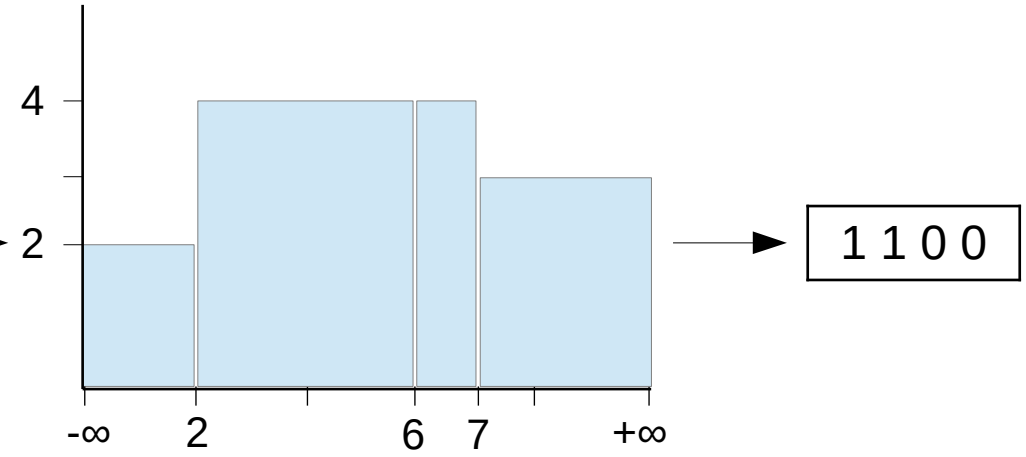
Querying

Range query: [1, 3]



Querying

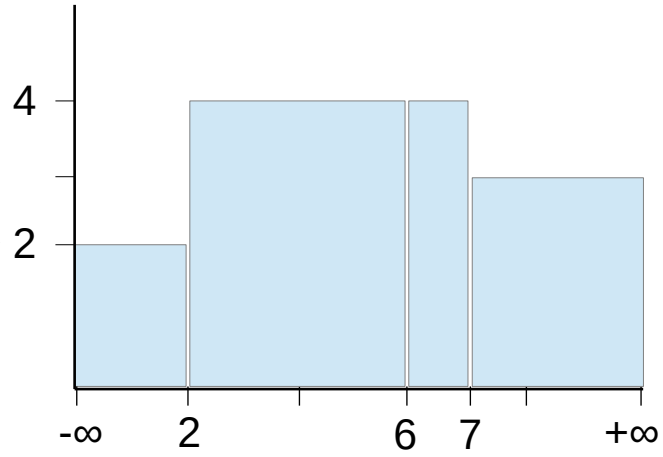
Range query: [1, 3]



Range query: [6, 8]

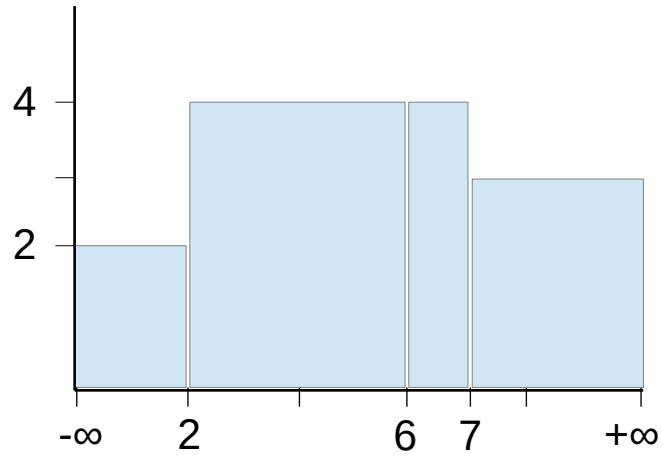
Querying

Range query: [1, 3]



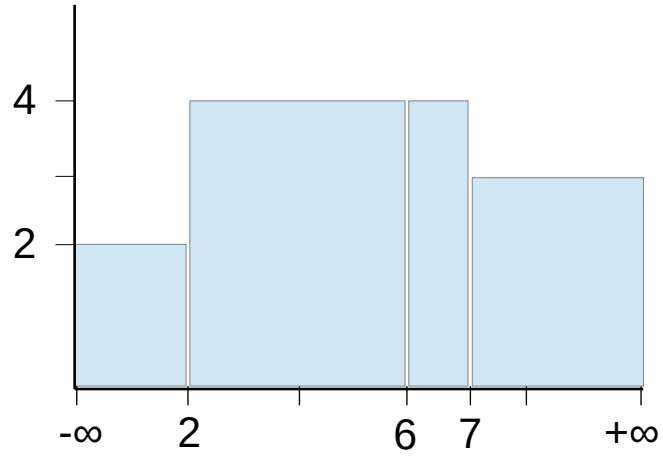
1 1 0 0

Range query: [6, 8]



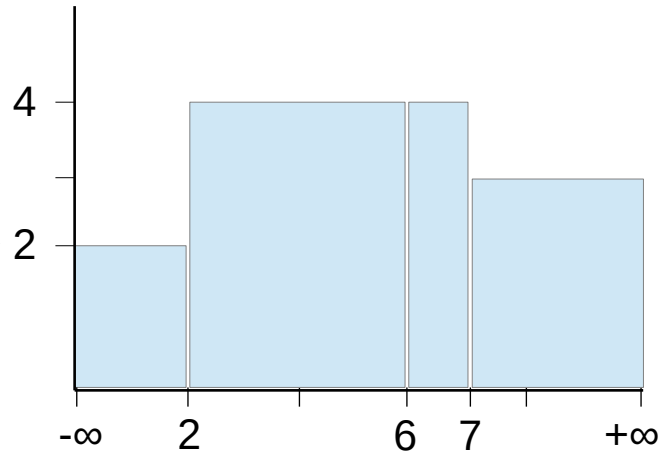
Querying

Range query: [1, 3]



1 1 0 0

Range query: [6, 8]



Mask

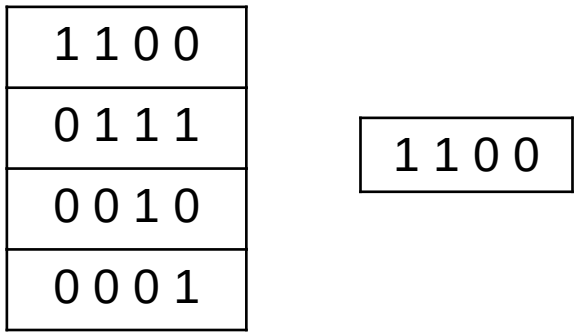
0 0 1 1

0 0 1 0

Innermask

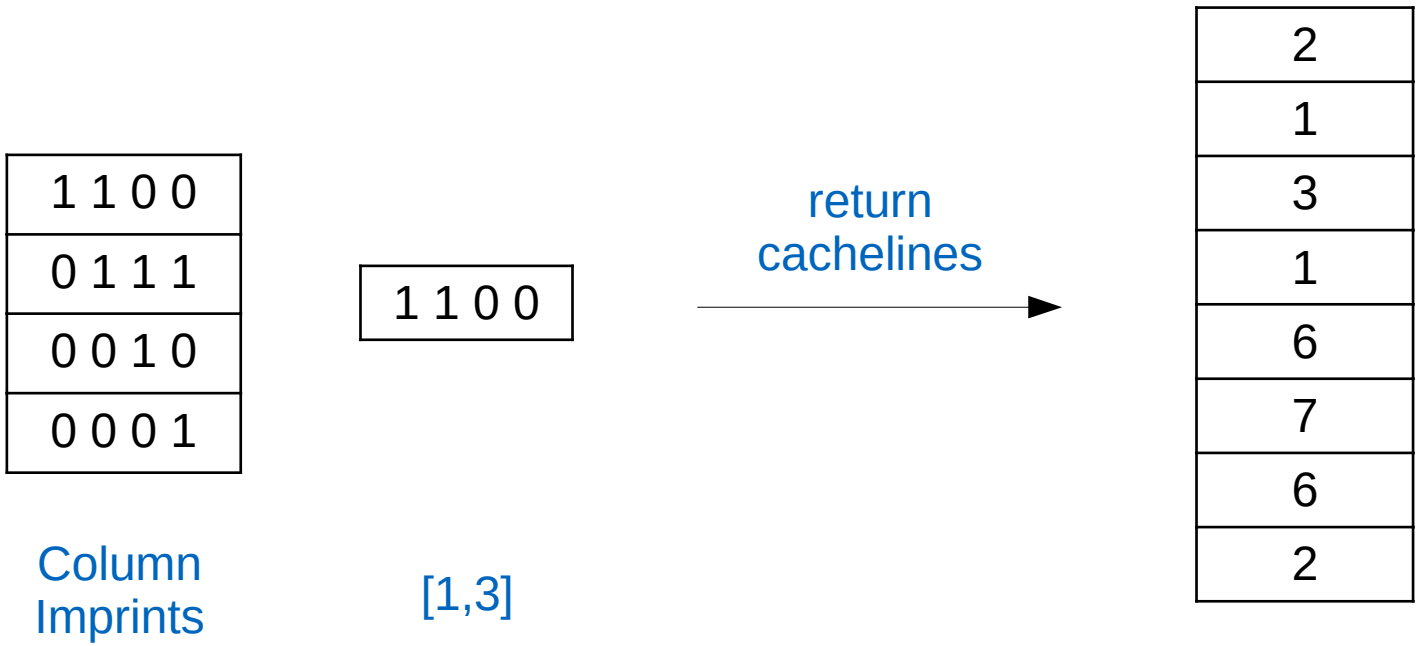
Querying

- Return cacheline if the query vector has common set bits with imprint
- If the imprint vector is exactly the same as the innermask, no need to check further



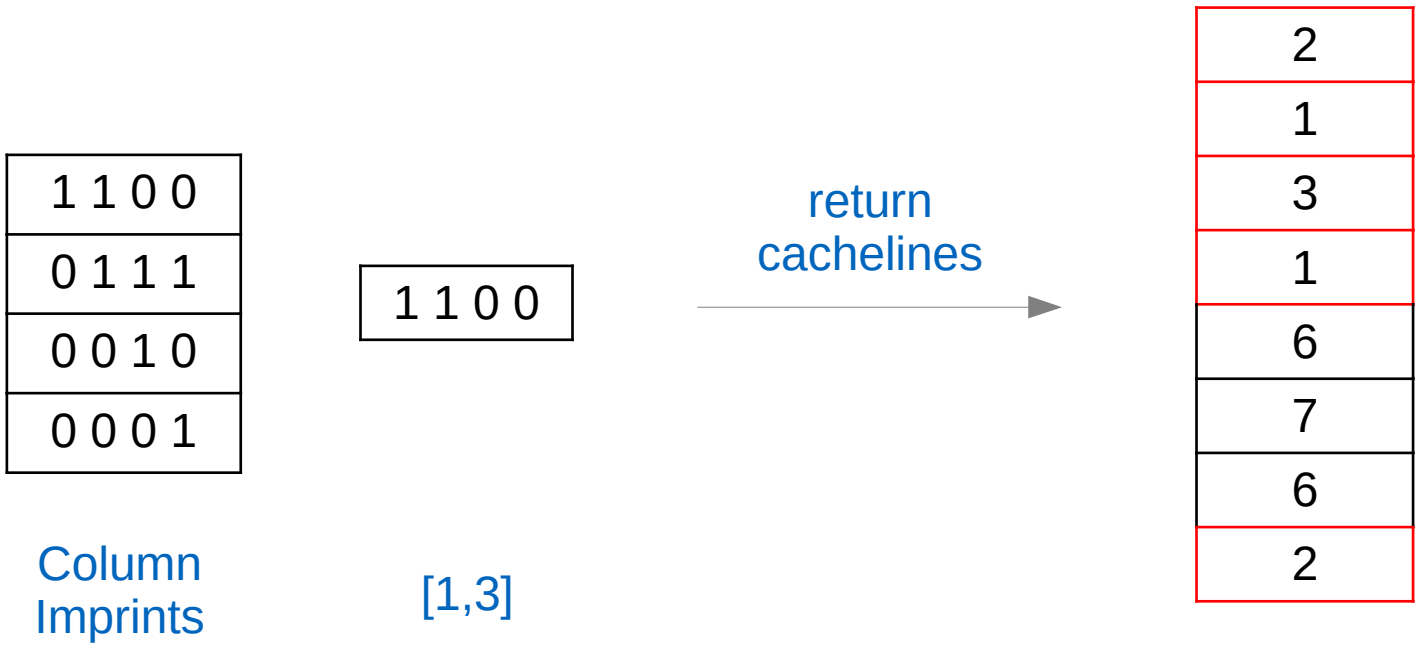
Querying

- Return cacheline if the query vector has common set bits with imprint
- If the imprint vector is exactly the same as the innermask, no need to check further



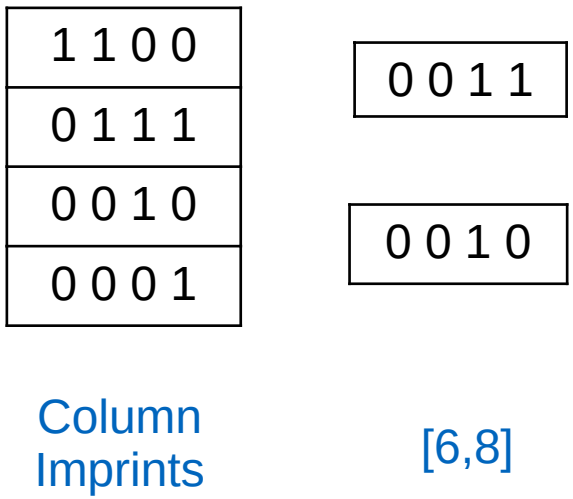
Querying

- Return cacheline if the query vector has common set bits with imprint
- If the imprint vector is exactly the same as the innermask, no need to check further



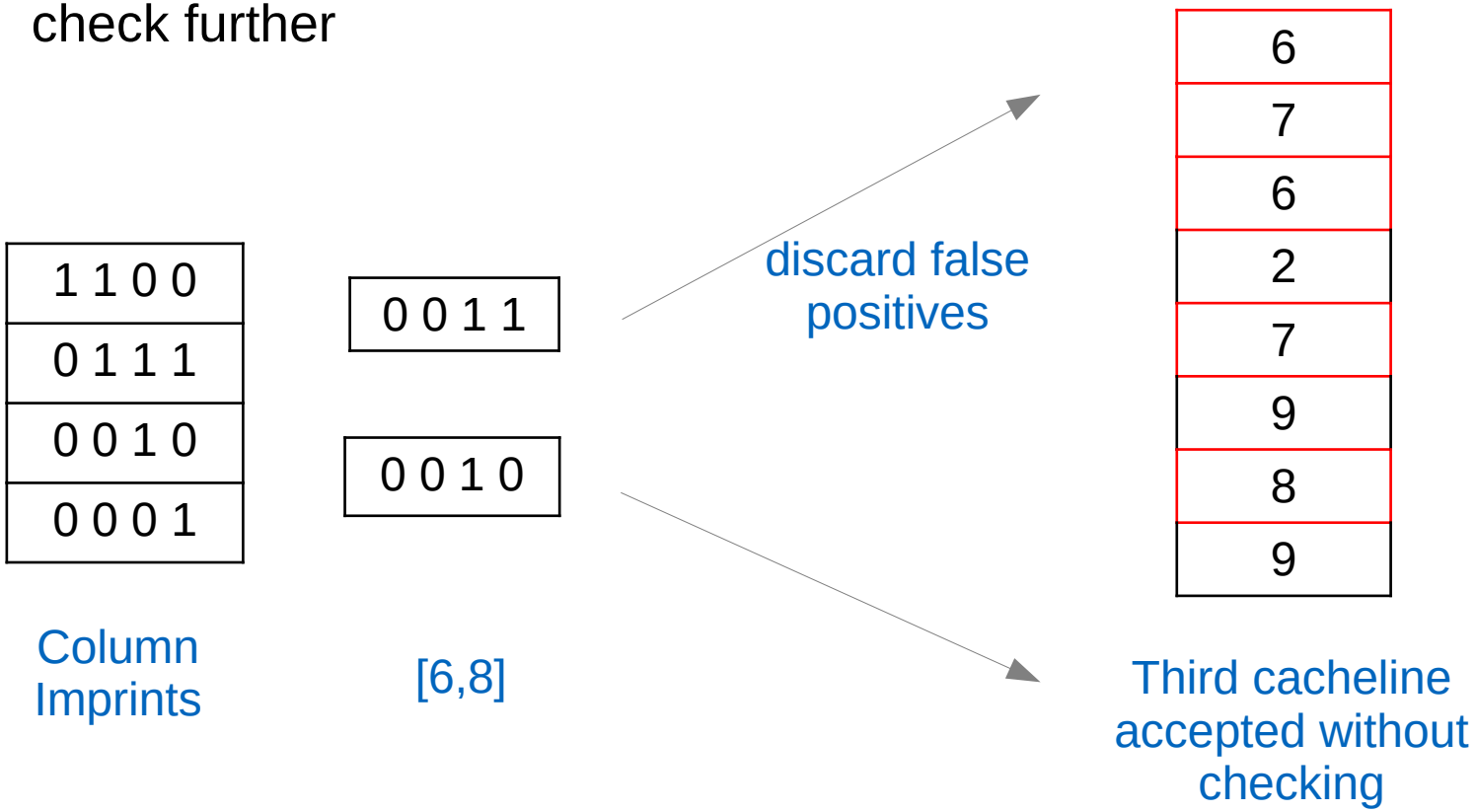
Querying

- Return cacheline if the query vector has common set bits with imprint
- If the imprint vector is exactly the same as the innermask, no need to check further

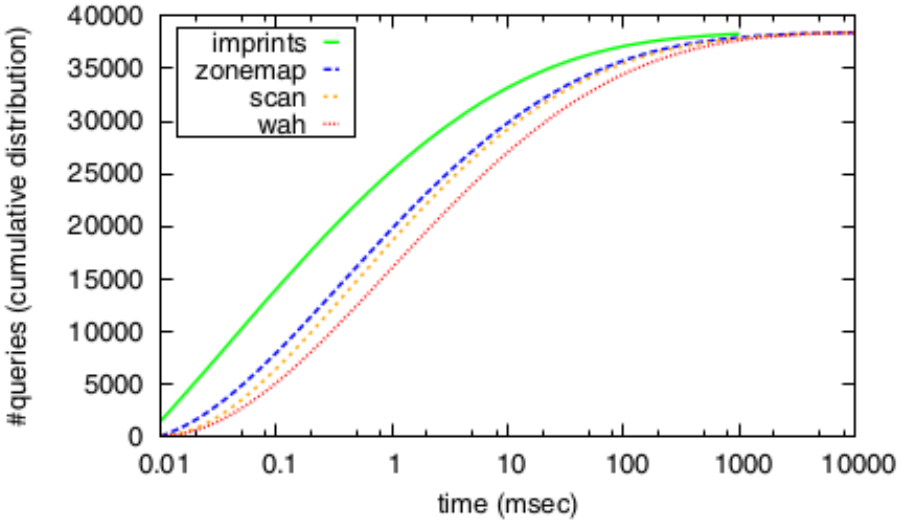


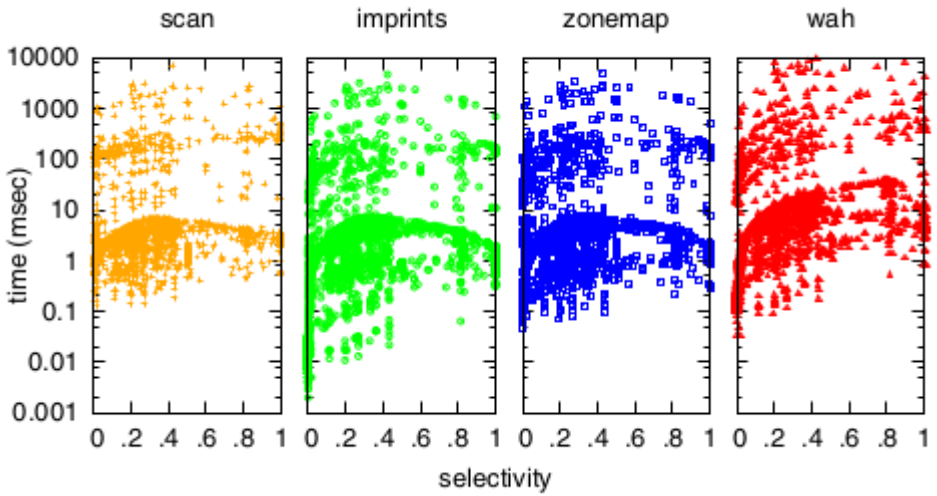
Querying

- Return cacheline if the query vector has common set bits with imprint
- If the imprint vector is exactly the same as the innermask, no need to check further



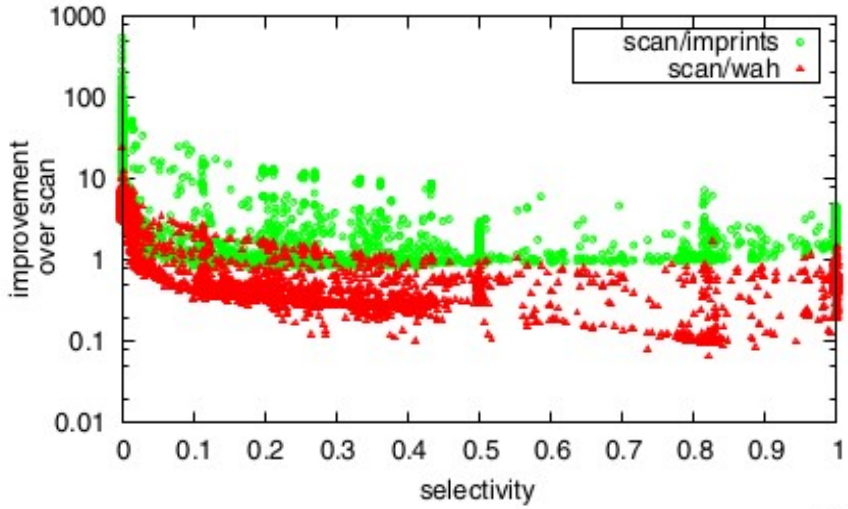
Results and benchmarks

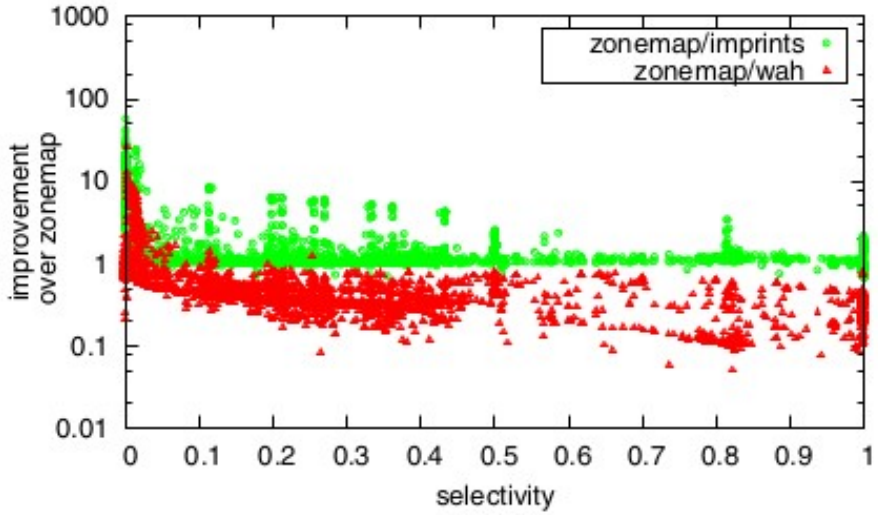




Cumulative distribution of query times and query time for decreasing selectivity [1]

Results and benchmarks





Factor improvement of imprints over scan and over zonemaps [1]

Summary

- Column imprints are used for faster search on locally clustered data
- They are presented as a space- and time-efficient solution
- They achieve better query times than scan and other index structures in high selectivity queries and with skewed data
- Their perfomance depends on how representative the histogram is

References

[1] Sidirourgos L., Kersten M. (2013). Column imprints: A secondary index structure. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data.*

[2] Hentschel B., Kester M. S., Idreos S. (2018), Column Sketches: A Scan Accelerator for Rapid and Robust Predicate Evaluation. *In ACM SIGMOD International Conference on Management of Data*.

Thank you!

Questions? :)