ETH zürich

Week 1 - SQL Brush-Up

Aljaž Medič

Plan for today

1. Short discussion about the course
2. Quiz
3. SQL

QUIZ

The three Vs of Big Data are Volume, Variety, and Velocity.

True.

Most companies delete their data as soon as it becomes outdated.

False. (Many companies keep a lot of their data for possible future use, unless prevented by regulations like GDPR.)

Giga is the standardized prefix for 10^12.

False. (**Tera** is the standardized prefix for 10^12; **Giga** is for 10^9.)

QUIZ

Name all five primary **data shape types** discussed for modern databases.

Tables, trees, cubes, graphs and text/vectors.

The technique that addresses the growing discrepancy between capacity and throughput is called **batch processing**.

False. (It's called parallelism.)

Data Independence refers to the ability to access physical storage details directly.

False. (Data independence means separating the logical view of data from its physical storage details.)

QUIZ

Data Totality means one must have complete data.

True.

A **Cartesian product** is the result of a projection.

False. (Cartesian product = all tuple combinations; projection = selecting columns.)

Durability in ACID ensures data survives crashes or power failures.

True.

QUIZ

Name all typical units used for capacity, throughput, and latency in data systems.

Capacity (bytes), throughput (bytes/second), latency (milliseconds)

The relational model enforces relational integrity, causational integrity, and atomic integrity.

False. (It enforces relational integrity, domain integrity, and atomic integrity.)

(atomic integrity is also known as the 1. NF)

Scaling up means distributing data across multiple machines.

False. (Scaling up = more powerful single machine; Scaling out = multiple machines.)

QUIZ

A relational database management system always exposes a logical model and building blocks for manipulating data, separate from the physical layer.

True.

Standardization of query languages enables query reuse across different database vendors.

True.

Relational tables contain nested data as their primary format.

False. (Relational tables are flat collections of records; nested data is not their primary format.)

QUIZ

What is the difference between WHERE and HAVING?

WHERE filters before grouping; HAVING filters after grouping

SQL is an imperative and functional language.

False. (It's declarative and functional.)

The SELECT clause in SQL performs a selection.

False. (It performs a projection.)

QUIZ

SQL is a set-based language.

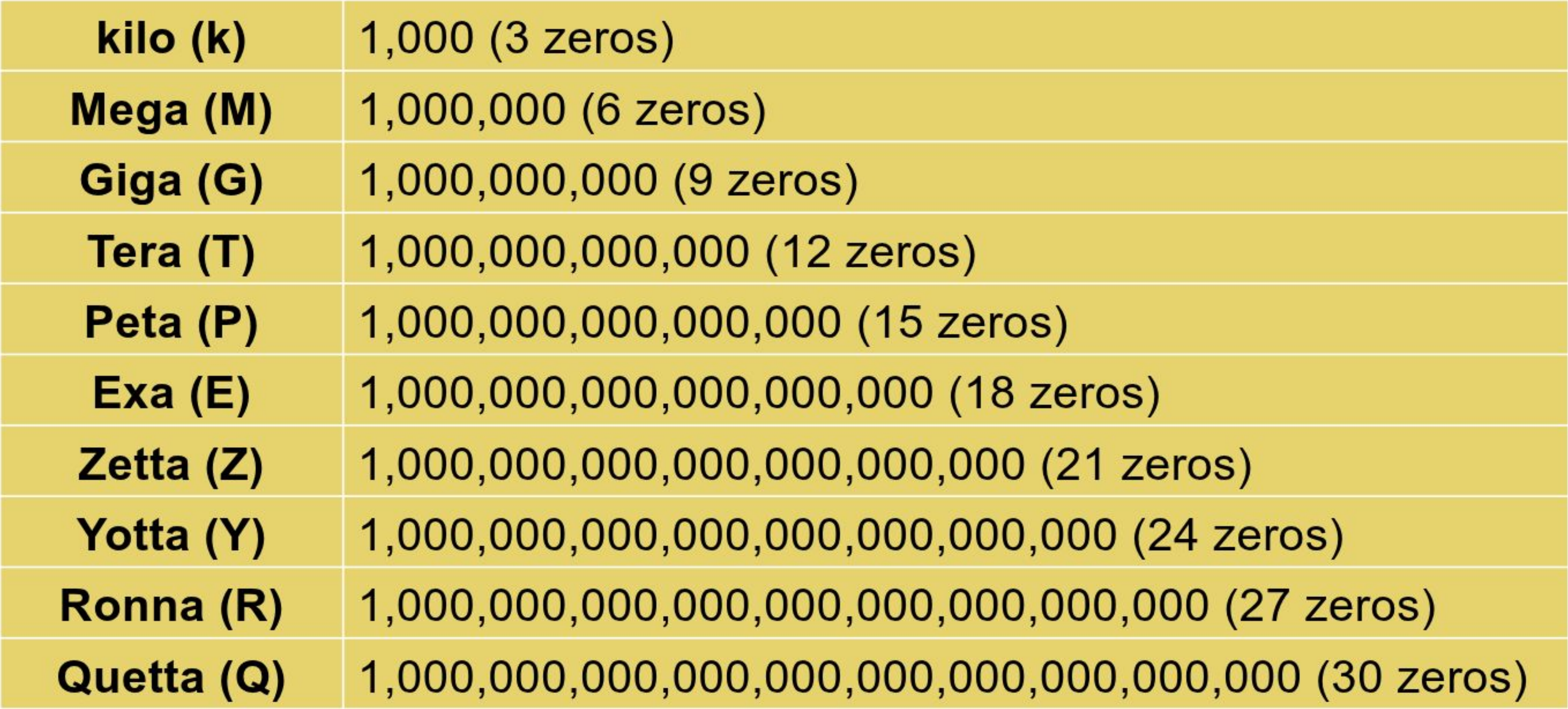
True.

List 3 types of atomic values typically considered in database systems.

Strings (!), Integers, Booleans, Dates,

etc.(all values not structured like objects, arrays, lists, or trees)

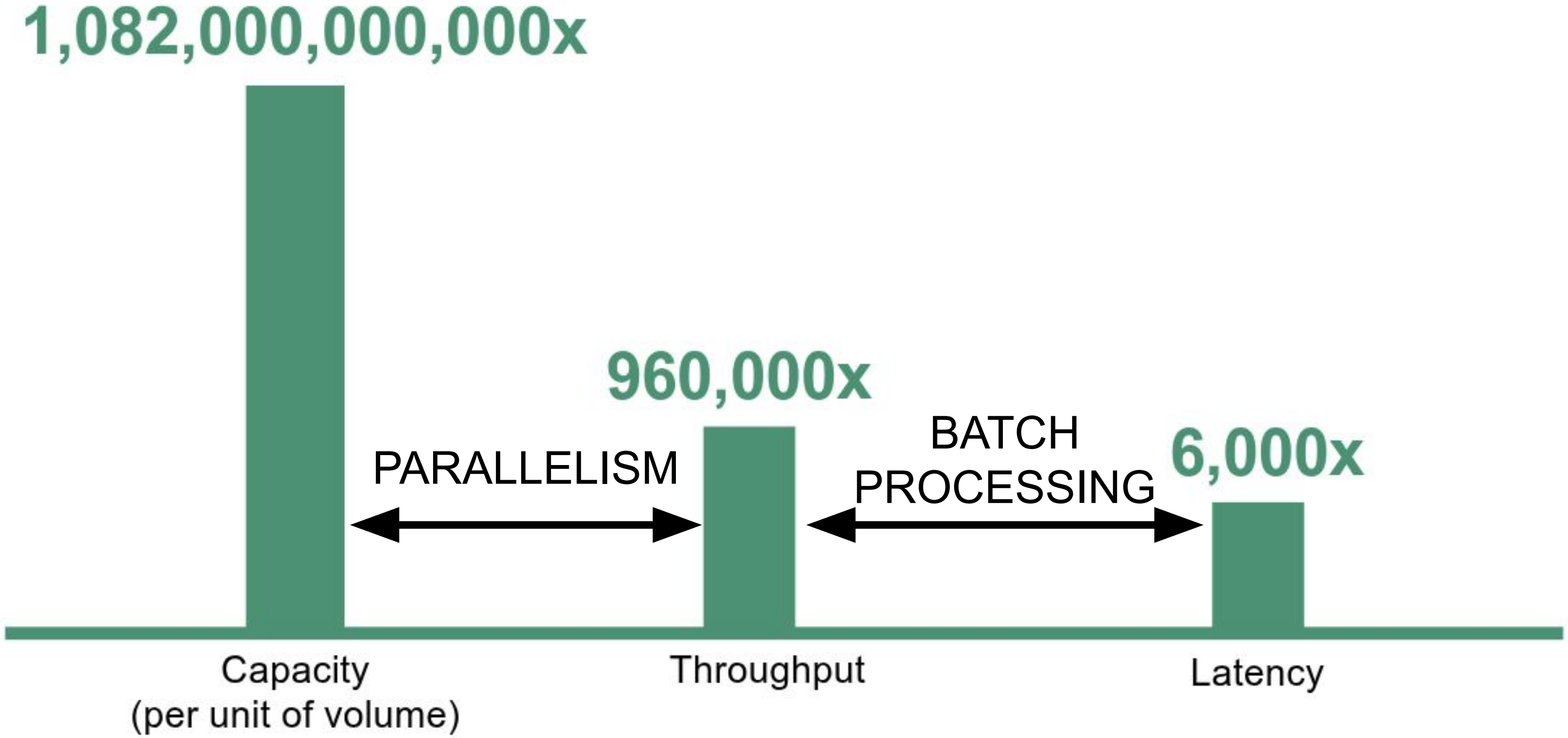
Prefixes



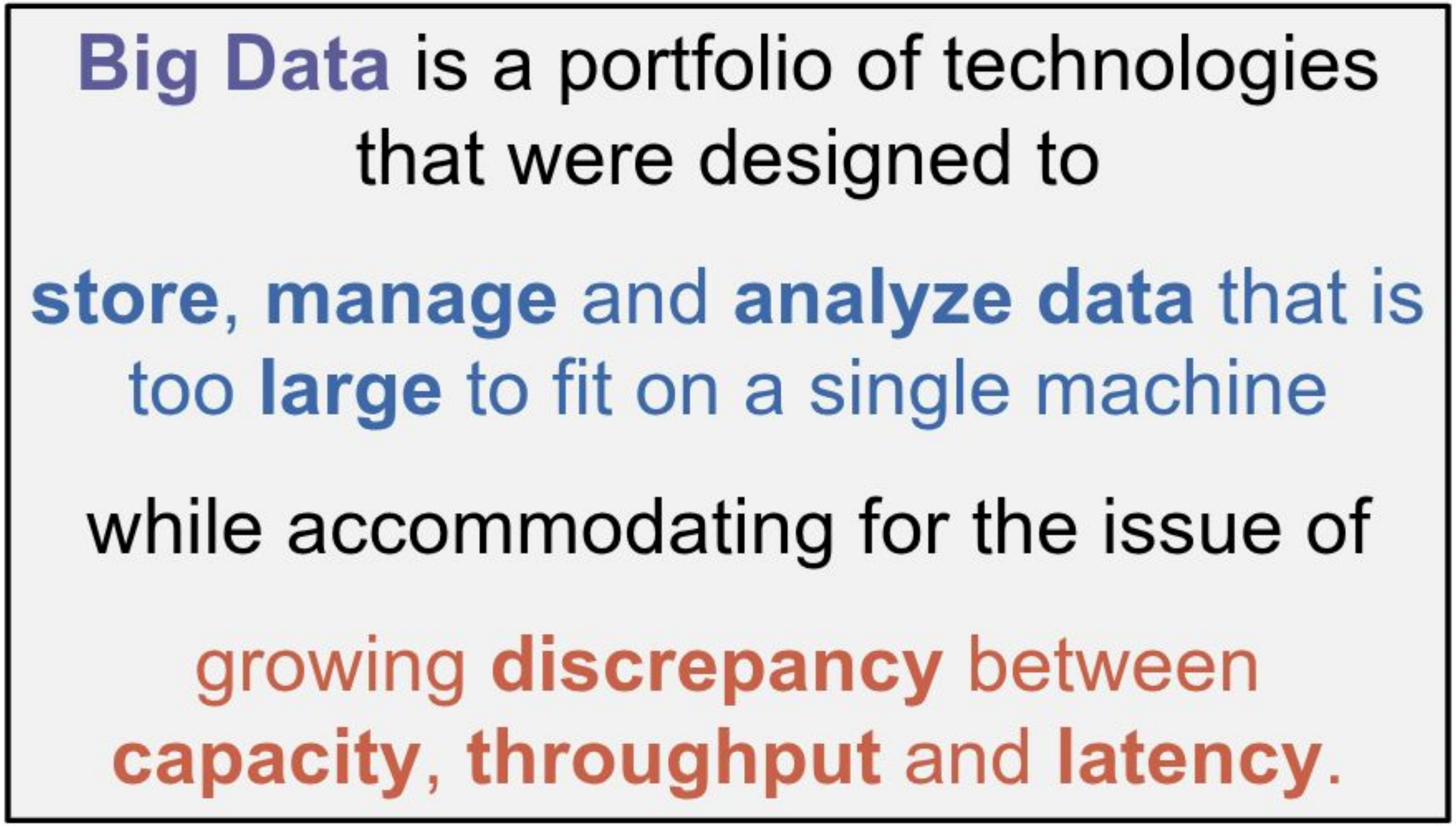
kilo (k)	1,000 (3 zeros)
Mega (M)	1,000,000 (6 zeros)
Giga (G)	1,000,000,000 (9 zeros)
Tera (T)	1,000,000,000,000 (12 zeros)
Peta (P)	1,000,000,000,000,000 (15 zeros)
Exa (E)	1,000,000,000,000,000,000 (18 zeros)
Zetta (Z)	1,000,000,000,000,000,000,000 (21 zeros)
Yotta (Y)	1,000,000,000,000,000,000,000,000 (24 zeros)
Ronna (R)	1,000,000,000,000,000,000,000,000,000 (27 zeros)
Quetta (Q)	1,000,000,000,000,000,000,000,000,000,000 (30 zeros)

Motivation

The progress made (1956-2025): Logarithmic



Definition from the Lectures



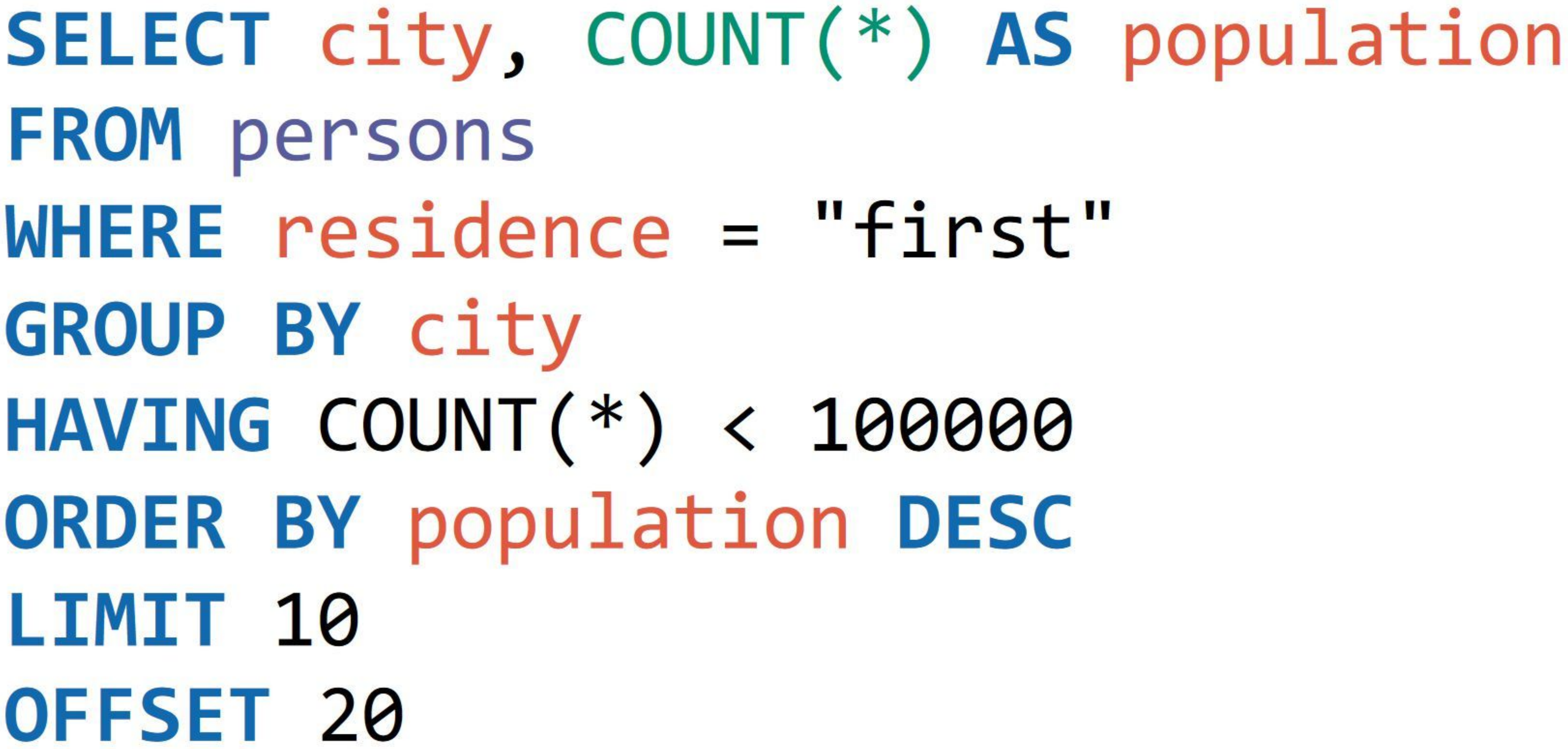
Big Data is a portfolio of technologies that were designed to

store, **manage** and **analyze data** that is too **large** to fit on a single machine

while accommodating for the issue of

growing **discrepancy** between **capacity**, **throughput** and **latency**.

All SQL Clauses



```
SELECT city, COUNT(*) AS population
FROM persons
WHERE residence = "first"
GROUP BY city
HAVING COUNT(*) < 100000
ORDER BY population DESC
LIMIT 10
OFFSET 20
```

Quick recap of **Relational Algebra**

Union
Intersection
Subtraction

Set queries

Selection
Projection

Filter queries

Relation renaming
Attribute renaming

Renaming queries

Grouping
Sorting

Shuffling queries

Cartesian product
Natural join
Theta join

Joining queries

Tips for solving the exercise sheet

- Use **CTEs** (Common Table Expression):

```
WITH cte_name AS (
    SELECT complex_query_to_be_reused
)
SELECT *
FROM cte_name;
```

- Rename tables for better readability:

```
SELECT t.id, t.name
FROM very_long_relation_name t;
```

- Usage of **NATURAL JOIN**
  (sometimes it’s better to do regular join + expression for more verbosity)

See you next week!

Aljaž Medič
amedic@ethz.ch

https://aljaz.si/teaching

Suggestions, Improvements