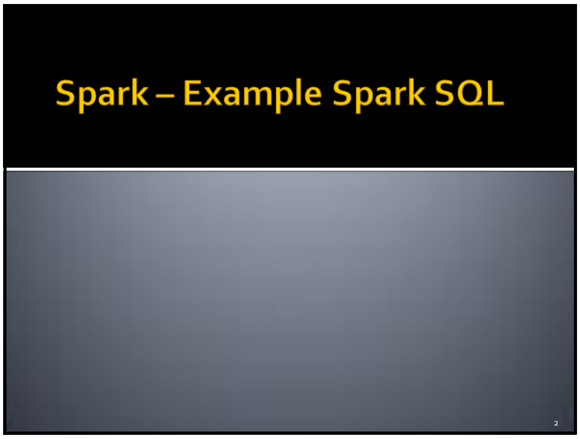
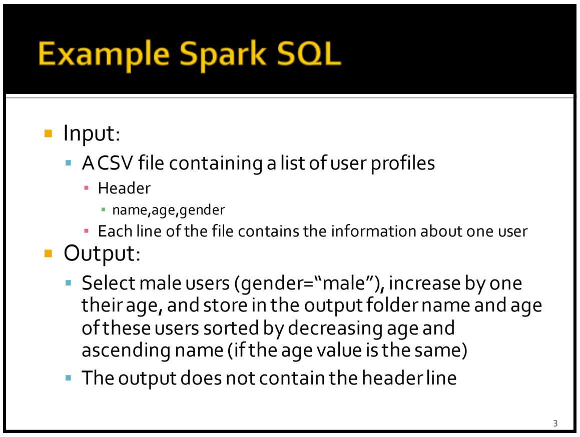
Big data: architectures and data analytics

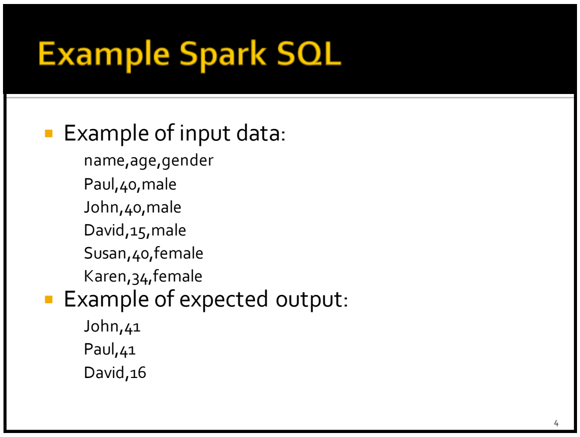


Spark – Example Spark SQL



Example Spark SQL

- Input:
  - A CSV file containing a list of user profiles
    - Header
      - name,age,gender
    - Each line of the file contains the information about one user
- Output:
  - Select male users (gender="male"), increase by one their age, and store in the output folder name and age of these users sorted by decreasing age and ascending name (if the age value is the same)
  - The output does not contain the header line



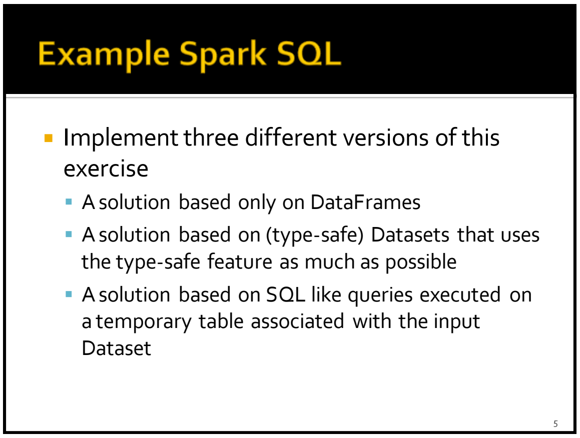
Example Spark SQL

- Example of input data:

```
name,age,gender
Paul,40,male
John,40,male
David,15,male
Susan,40,female
Karen,34,female
```

- Example of expected output:

```
John,41
Paul,41
David,16
```



Example Spark SQL

- Implement three different versions of this exercise
  - A solution based only on DataFrames
  - A solution based on (type-safe) Datasets that uses the type-safe feature as much as possible
  - A solution based on SQL like queries executed on a temporary table associated with the input Dataset