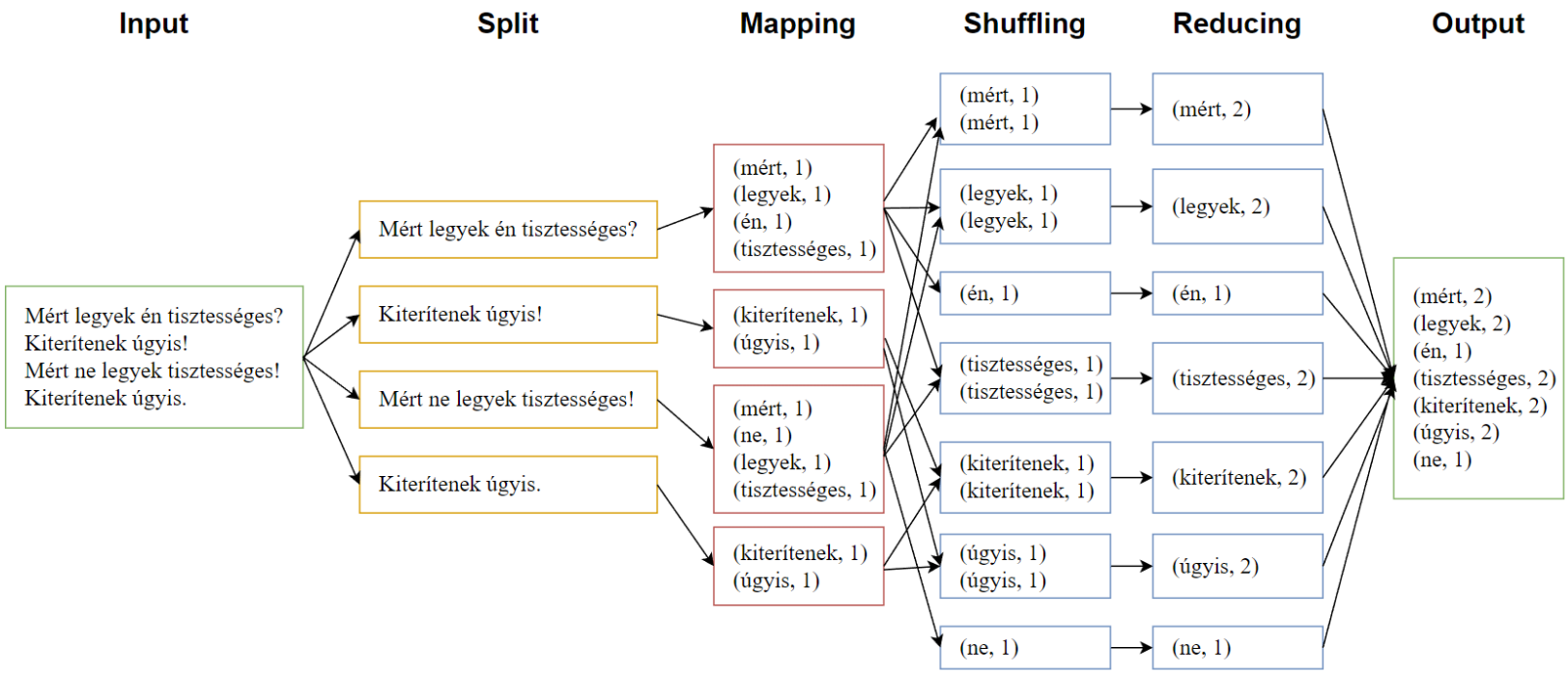
ELTE | IK

INFORMATIKAI KAR

BigData architektúrák és elemző módszerek

2. gyakorlat

Feladat: WordCount



Szöveg: József Attila - Két hexameter

1. részfeladat

- Írjuk meg a WordCount Mapper osztályát.
- Hozzunk létre több bemeneti fájlt és vizsgáljuk meg a kimeneteket!
- A teszteléshez állítsuk be azt, hogy nincs szükségünk Reducer osztályra:

```
job.setNumReduceTasks(0);
```

Hasznos megjegyzés

- Ha nem szeretnénk a kimeneti mappát minden futtatásnál kézzel törölni írjuk az alábbi kódrészletet a Driver osztály main() függvényébe.
- Az outPath az a kimeneti útvonalunk legyen.

```
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;

...

Path outPath = new Path("out");
FileOutputFormat.setOutputPath(job, outPath);

FileSystem fs;
fs = FileSystem.get(conf);
if(fs.exists(outPath)) {
    fs.delete(outPath, true);
}
```

2. részfeladat

- Írjuk meg a WordCount Reducer osztályát.
- A Reducer egy összegzést végezzen a kapott kulcs-érték párokon.

3. részfeladat

- Használjunk Combiner osztályt a szavak lokális aggregációjára.
- Ehhez állítsuk be a korábban elkészített Reducer osztályt combinerként a driverben:

```
job.setCombinerClass(WordCountReducer.class);
```

Feladatok a WordCount továbbfejlesztéséhez

1. A Mapper osztályban távolítsuk el a szavak végéről a speciális karaktereket (',', '!', '?', ':' stb.)
2. Alakítsuk kisbetűssé a szavakat.
3. Szűrjük ki a legfeljebb 2 karakter hosszú szavakat.
4. Futtassuk a kódunkat valós bemeneti adatra (pl. újság cikk).

Feladat: Grep

- Készítsünk egy MapReduce programot, amely egy karaktersorozatot keres a bemeneti fájlban. Írjuk ki azokat a sorokat a bementi fájlból, amelyek tartalmazzák a keresett karaktersorozatot.
- Pl:
  - Keresett karaktersorozat: „sütsz”
  - Bemeneti fájl tartalma:

    Mit sütsz kis szűcs,

    tán sós húst sütsz

    kis szűcs?
  - Kimeneti fájl elvárt tartalma:

    Mit sütsz kis szűcs,

    tán sós húst sütsz

Feladat: Grep

- A grep feladat esetében a reducer akár el is hagyható, az alábbi sor driver-be írásával:

```
job.setNumReduceTasks(0);
```

Feladat: WordMean

- Adjuk meg, hogy a bemeneti fájlban mekkora a szavak átlagos hossza
- Példa bemenet: „Tedd, vagy ne tedd, de ne próbáld!”
- Példa kimenet: 3.57