**Exercise 1:**

**Learning goal:**

* Read/Write data from HDFS
* JOIN
* FILTER
* Aggregate
* Remove duplicates from a dataset

**Business case:** You would need to help a transport company to prepare following details for better insight into their business

1. List of all the drivers who made at least one mistake (Events other than Normal)
2. Would like to know if there is any driver who is not certified even though driving
3. Store each driver’s aggregate work history in following format:

id, name, wage-plan, total\_hours, total\_miles

**Reference Commands**

* Load without alias

truck\_events = LOAD 'truck\_event\_text\_partition.csv' USING PigStorage(',');

* DESCRIBE a Relation

DESCRIBE truck\_events;

* DUMP data on console

DUMP truck\_events;

* Load with alias

truck\_events = LOAD 'truck\_event\_text\_partition.csv' USING PigStorage(',') AS (driverId:int, truckId:int, eventTime:chararray, eventType:chararray, longitude:double, latitude:double, eventKey:chararray, correlationId:long, driverName:chararray, routeId:long, routeName:chararray, eventDate:chararray);

DESCRIBE truck\_events;

* Take sample records

truck\_events\_subset = LIMIT truck\_events 100;

dump truck\_event\_subset;

* Choose specific columns

specific\_columns = FOREACH truck\_events\_subset GENERATE driverId, eventTime, eventType;

DESCRIBE specific\_columns;

DUMP specific\_columns;

* STORE output

STORE specific\_columns INTO 'output\_directory' USING PigStorage(',');

* JOIN two datasets

truck\_events = LOAD 'truck\_event\_text\_partition.csv' USING PigStorage(',')

AS (driverId:int, truckId:int, eventTime:chararray,

eventType:chararray, longitude:double, latitude:double,

eventKey:chararray, correlationId:long, driverName:chararray,

routeId:long,routeName:chararray,eventDate:chararray);

drivers = LOAD 'drivers.csv' USING PigStorage(',')

AS (driverId:int, name:chararray, ssn:chararray,

location:chararray, certified:chararray, wage\_plan:chararray);

join\_data = JOIN truck\_events BY (driverId), drivers BY (driverId);

DESCRIBE join\_data;

DUMP join\_data;

* SORT

ordered\_data = ORDER drivers BY name asc;

DUMP ordered\_data;

* FILTER

filtered\_events = FILTER truck\_events BY NOT (eventType MATCHES 'Normal');

DUMP filtered\_events;

* SPLIT FILTER

SPLIT filtered\_events INTO normal\_events if eventType == 'Normal', others if eventType != 'Normal';

dump normal\_events;

dump others;

* GROUP

grouped\_events = GROUP filtered\_events BY driverId;

DESCRIBE grouped\_events;

DUMP grouped\_events;