Institute of Computer Technology

B. Tech. Computer Science and Engineering

Sub: Big Data Analytics

Project Report

1. AIM of the project given to you.

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management.

2. Name of students along with enrolment numbers who were included in the project- along with contribution of each student in the project.

Pandey Manthan (19162121016)
Parmar Ankit (19162121017)
Patel Darshil (19162121018)

3. The tool chosen to execute the project in- Explanation of the tool as well as the conceptualization of using the tool WRT the project.

What is hive?

Hive in Big Data is an easy-to-use software application that lets one analyze large-scale data through the batch processing technique. An efficient program, it uses a familiar software that uses HiveQL, a language that is very similar to SQL- structured query language used for interaction with databases.

Such a software can be operated by both programmers and non-programmers, making it a very accessible and easy-to-use application for converting petabytes of data into useful data strands. This is one of the biggest benefits of Apache Hive that has made it a popular choice for data analytics among large organizations with vast data.

Why are we using hive?

Hive was created to allow non-programmers familiar with SQL to work with petabytes of data, using a SQL-like interface called HiveQL. Traditional relational databases are designed for interactive queries on small to medium datasets and do not process huge datasets well. Hive instead uses batch processing so that it works quickly across a very large distributed database.

Hive transforms HiveQL queries into MapReduce or Tez jobs that run on Apache Hadoop's distributed job scheduling framework, Yet Another Resource Negotiator (YARN). It queries data stored in a distributed storage solution, like the Hadoop Distributed File System. Hive stores its database and table metadata in a metastore, which is a database or file backed store that enables easy data abstraction and discovery.

4. Code used for execution of the project.

1. Load csv file:

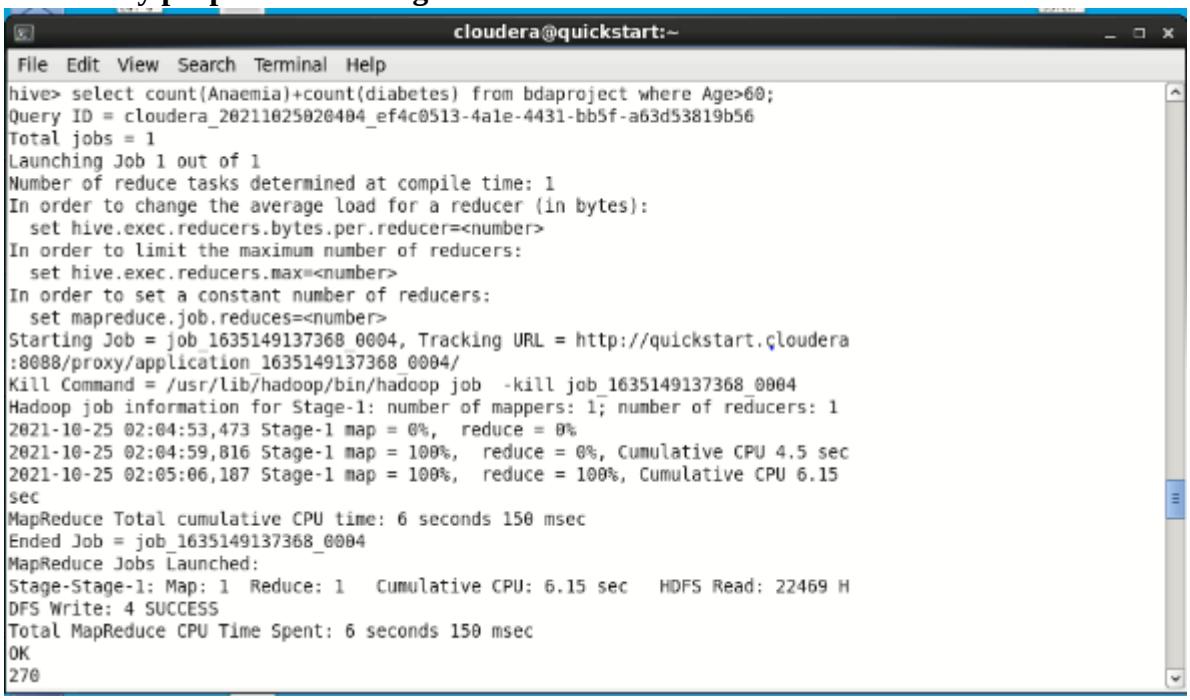
- cd Desktop
- hadoop fs -mkdir miniproject
- Hadoop fs -put heart_failure_clinical_records_dataset.csv miniproject
- Hadoop fs -ls miniproject
- hive
- create table bda_project
- (age INT, anaemia SMALLINT, creatinine_phosphokinase INT, diabetes SMALLINT, ejection_fraction INT, high_blood_pressure SMALLINT, platelets INT, serum_creatinine FLOAT, serum_sodium INT, sex BINARY, smoking SMALLINT, time INT, DEATH_EVENT SMALLINT)
- row format delimited fields terminated by ',';
- LOAD DATA LOCAL INPATH 'Desktop/ heart_failure_clinical_records_dataset.csv' INTO TABLE bda_project
- select *from bda_project;

2. Execution of project:

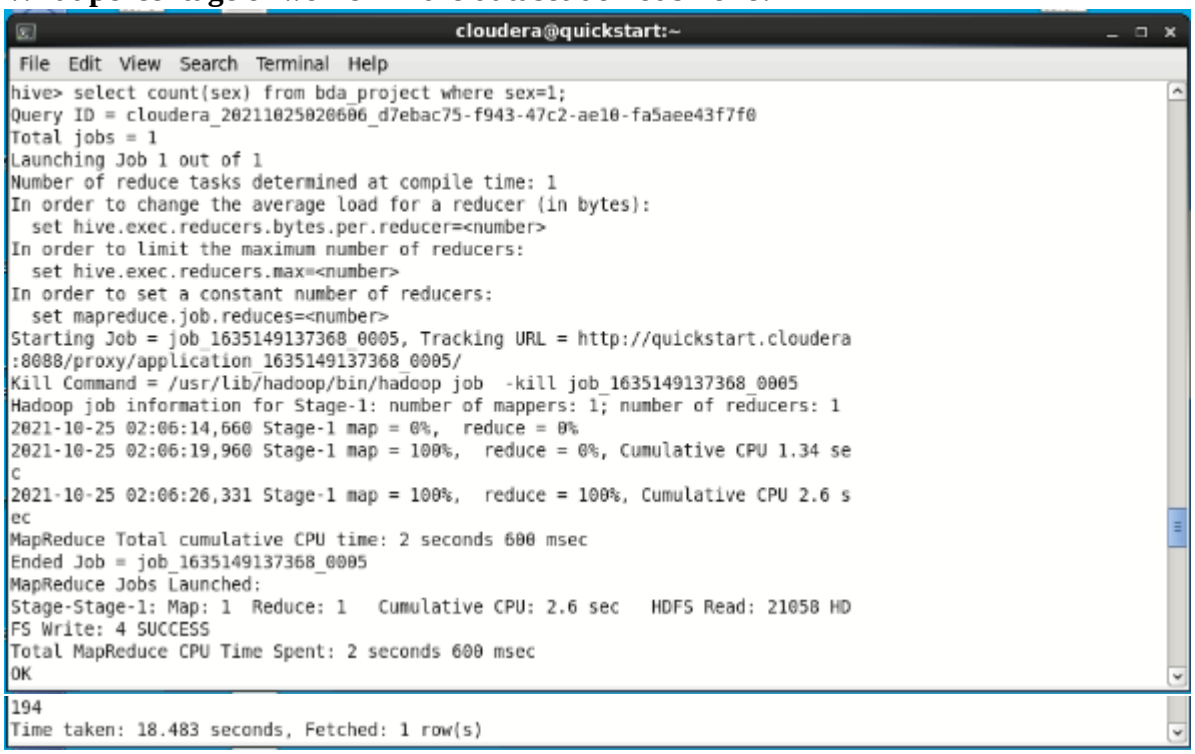
- select count(anaemia)+count(diabetes) from bda_project where Age>60;
- select count(sex) from bda_project where sex=1;
- select (count(sex)/194)*100 from bda_project where smoking=0 AND sex=1;
- select count(sex) from bda_project where age<50 AND DEATH_EVENT=1;
- select count(sex) from bda_project where serum_creatinine > 2 AND sex = 1;
- select count(sex) from bda_project where sex = 0 and time <= 20;
- select count(sex) from bda_project;
- select (count(sex)/299)*100 from bda_project where (anaemia=0 OR diabetes=0) AND smoking=0 AND DEATH_EVENT=1;
- select count(sex) from bda_project;
- select (count(sex)/299)*100 from bda_project where high_blood_pressure=1 AND DEATH_EVENT=0;
- select count(sex) from bda_project where platelets < 150000;
- select count(sex) from bda_project where high_blood_pressure=1;

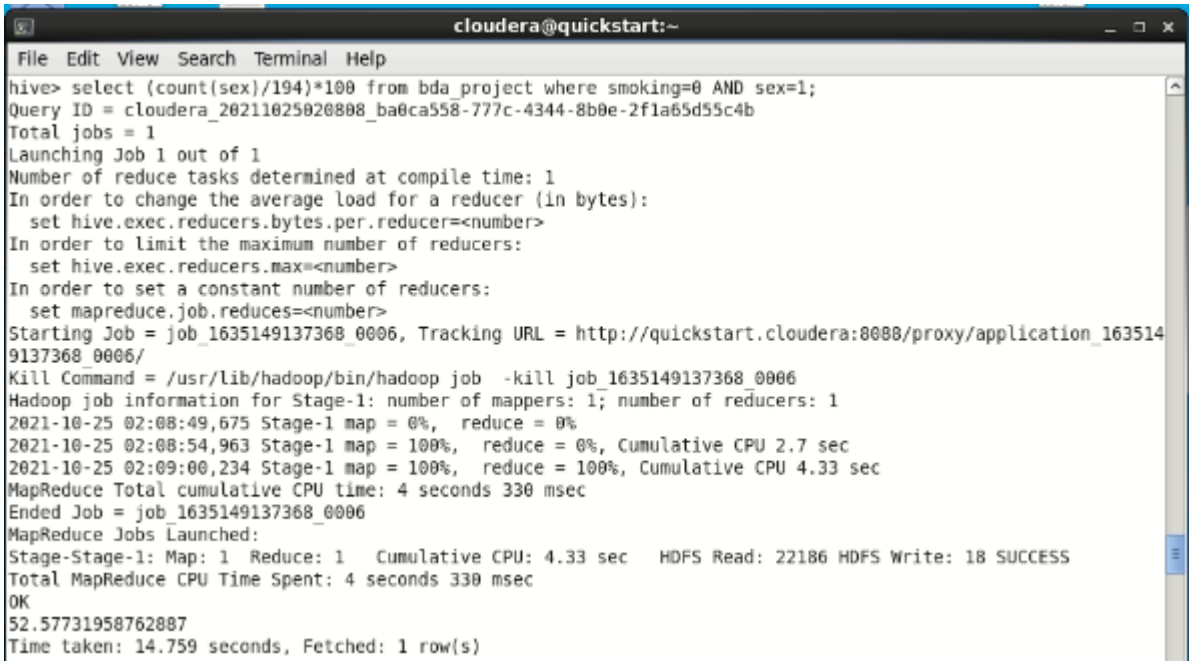
5. Problem definition and output:

- **How many people above the age of 60 are both anaemic and diabetic?**

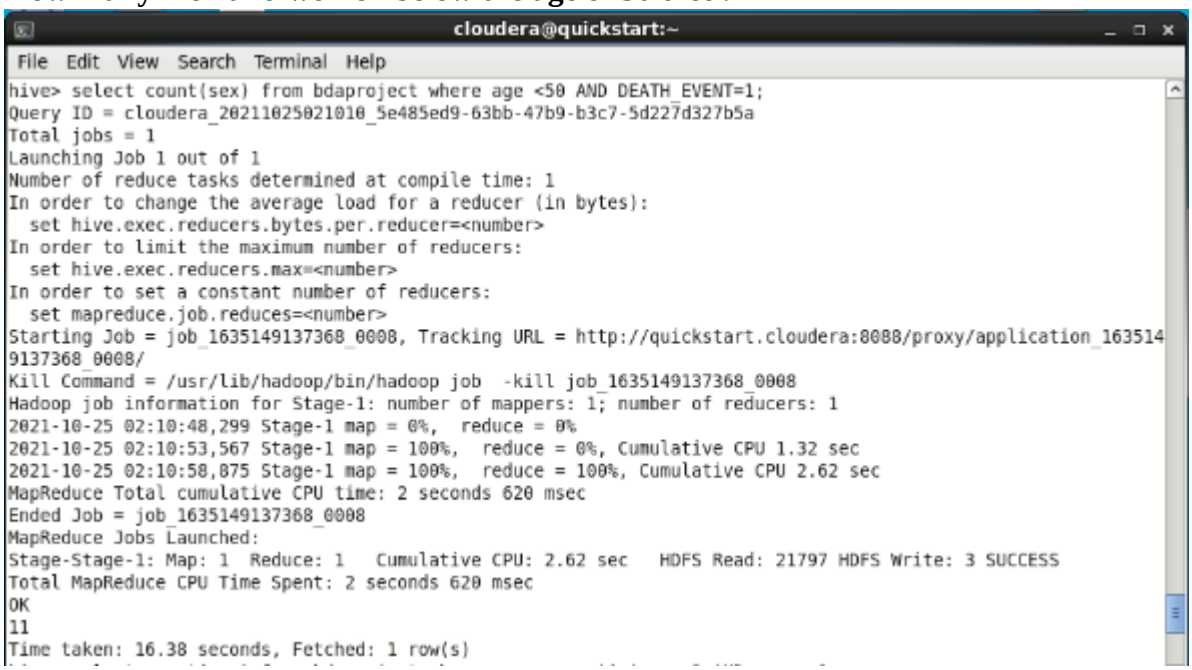


- **What percentage of women in the dataset do not smoke?**

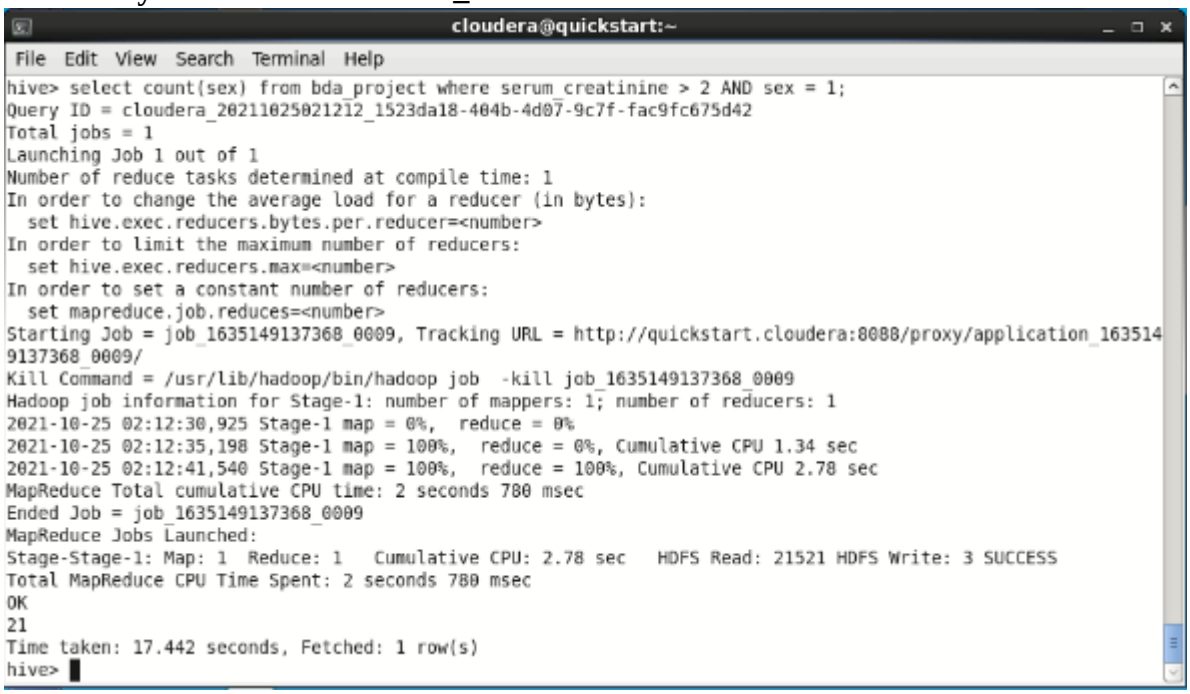




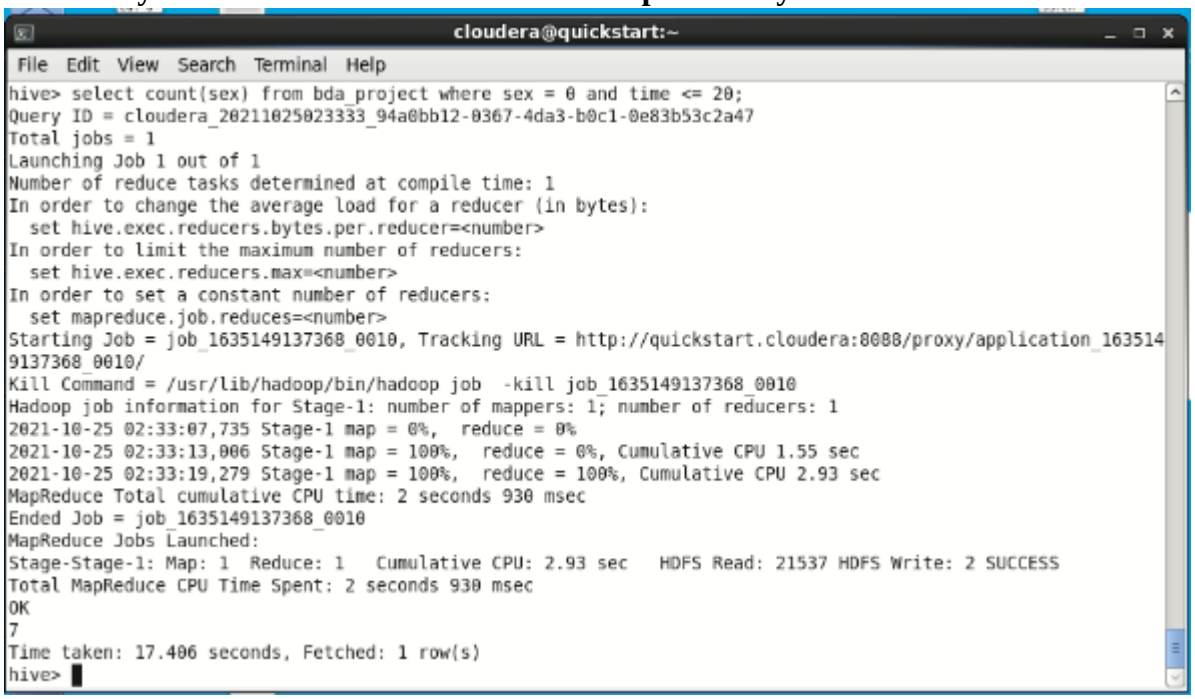
- **How many men and women below the age of 50 died?**



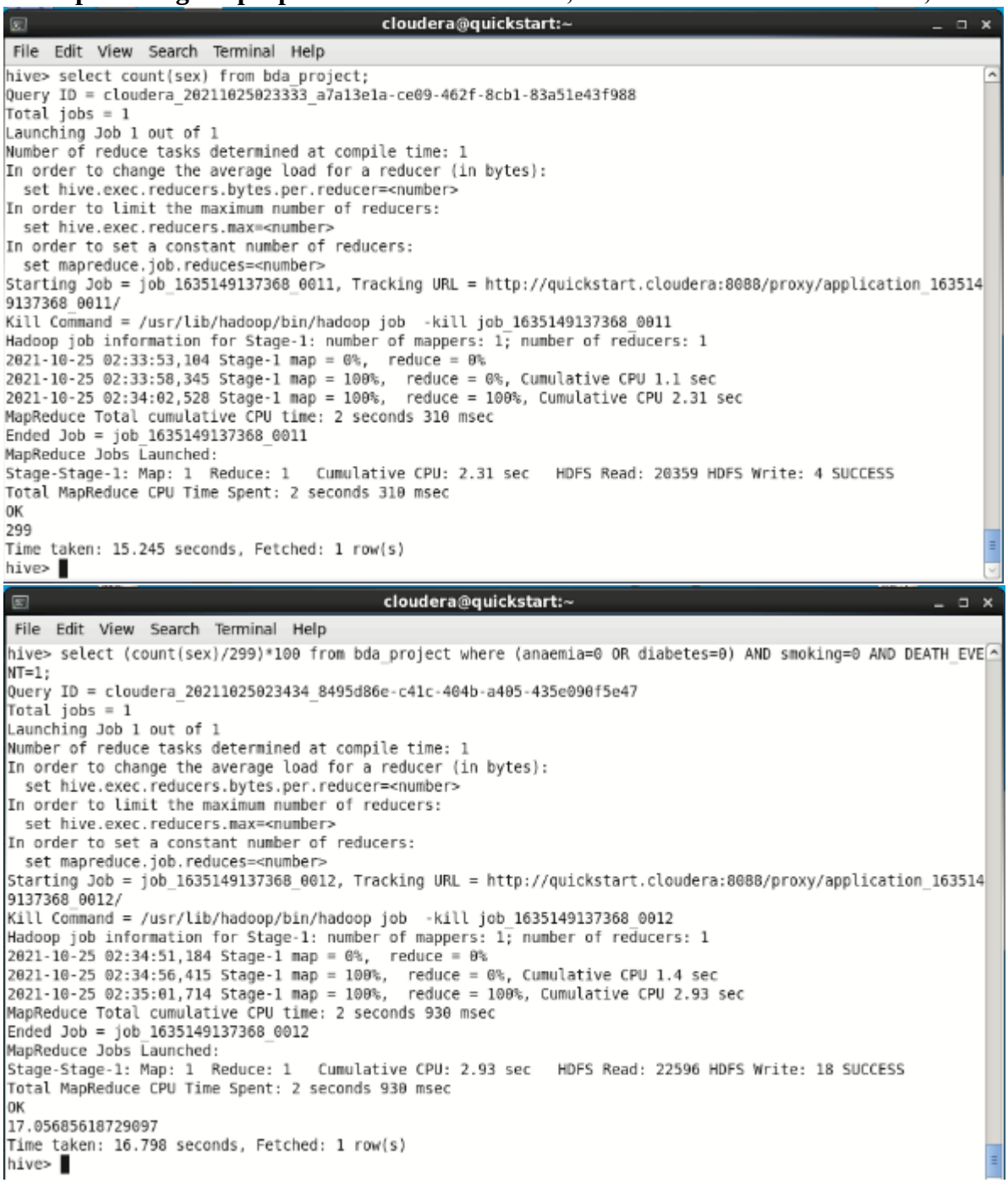
- **How many females have a serum_creatinine level of above 2?**



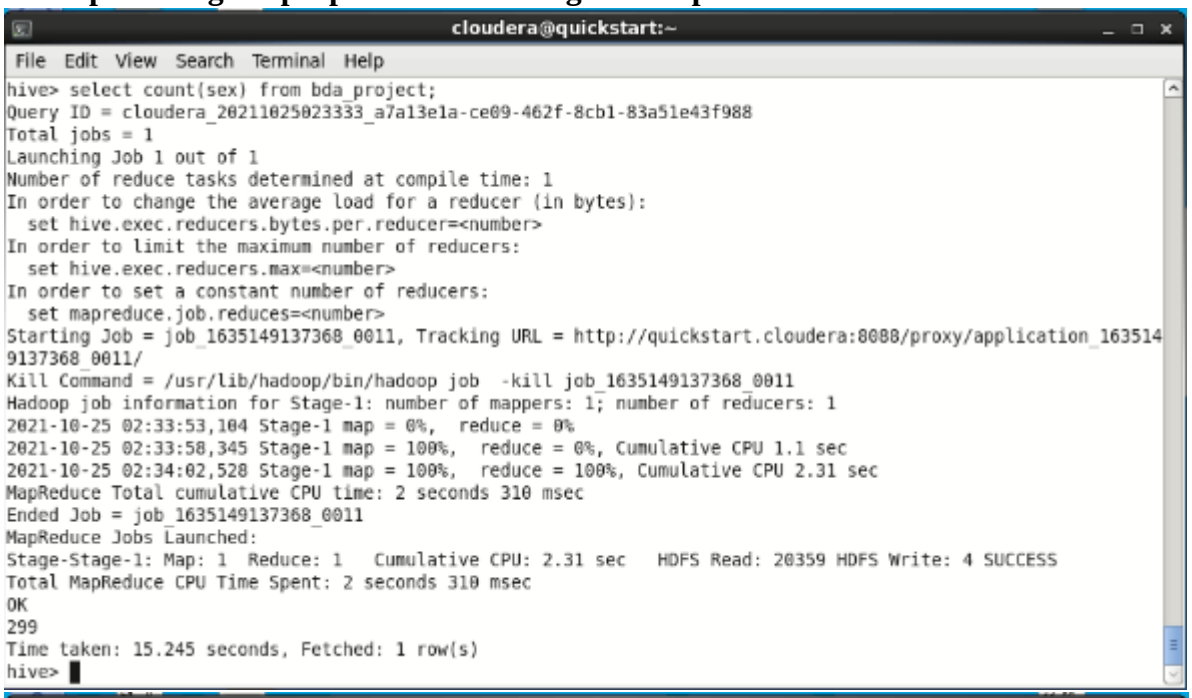
- **How many men have been called for a follow up in 20 days?**

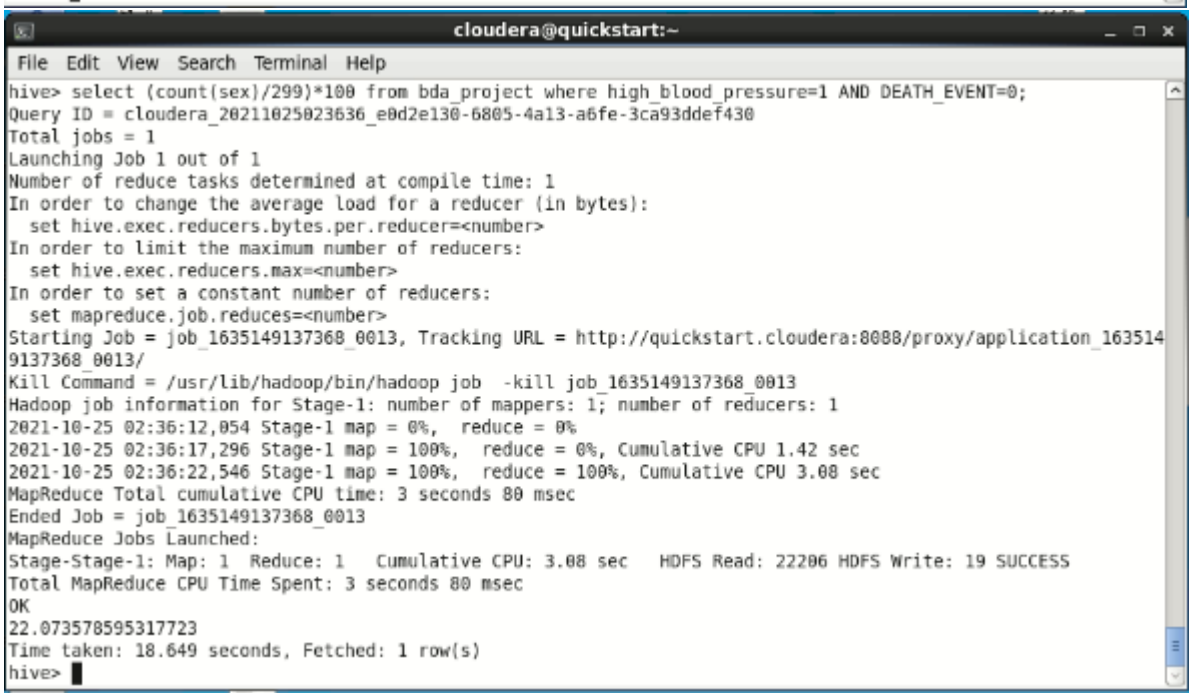


- **What percentage of people who weren't anaemic, or diabetic and did not smoke, died?**

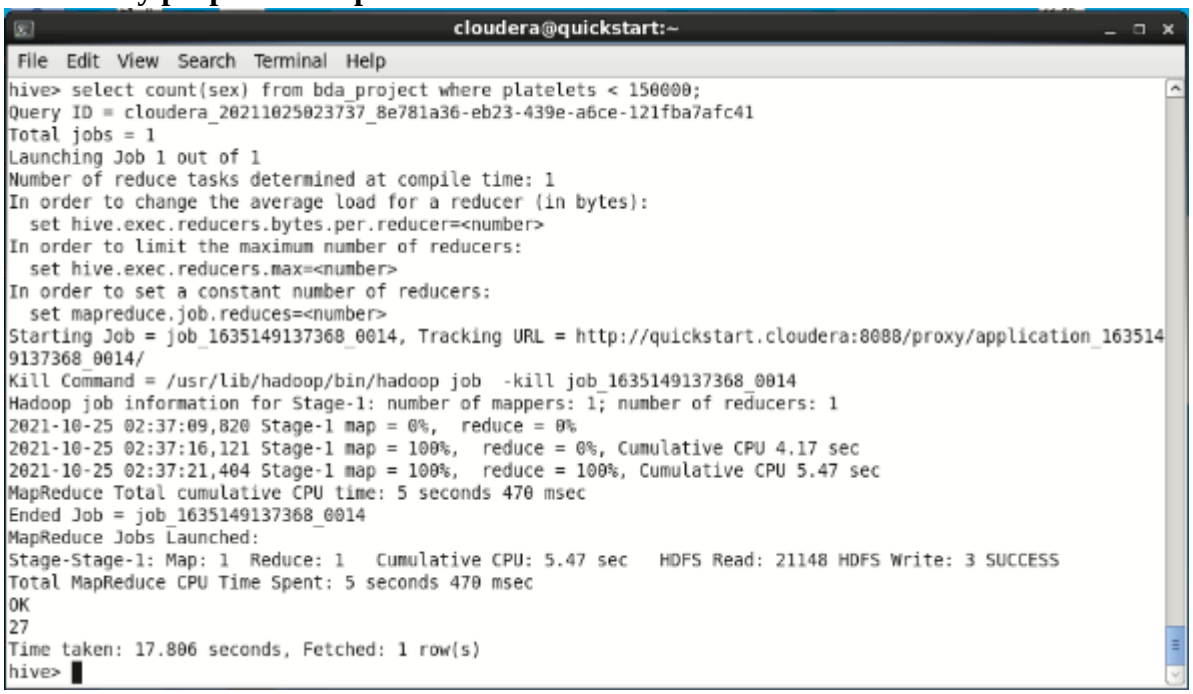


- **What percentage of people who have a high blood pressure did not die?**

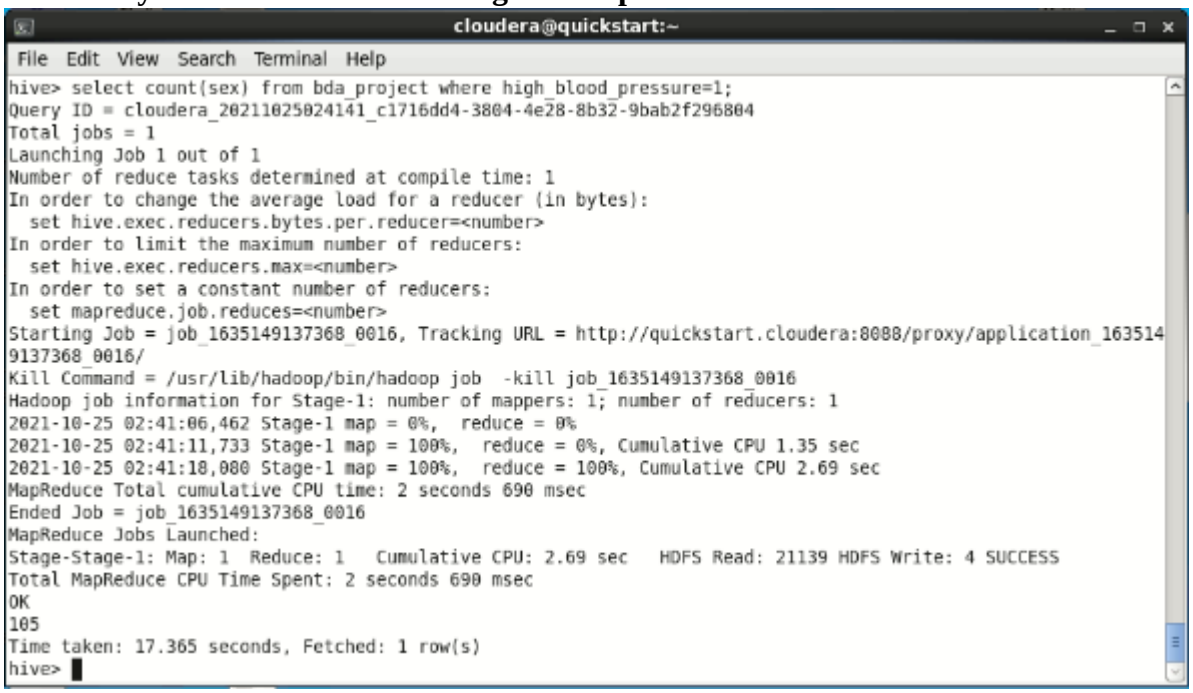




- **How many people have a platelet count of lesser than 150000?**



- **How many men and women have high blood pressure?**



6. Conclusion

Hive is a Data Warehousing package built on top of Hadoop used for data analysis. Hive also uses a language called HiveQL (HQL) which automatically translates SQL-like queries into MapReduce jobs. We have also learned various components of Hive like meta store etc.