Q1.a

create "starwarschars" , "cf"

org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.columns=HBASE\_ROW\_KEY,name,height ,mass,hair\_color, skin\_color ,eye\_color,birth\_year,gender ,homeworld ,species -Dimporttsv.separator=',' starwarschars /home/danish/Desktop/starwars/data

Q1.b

import pandas as pd

import os

pathdata = '/home/danish/Desktop/starwars/data'

os.chdir(pathdata)

os.getcwd()

starchars = pd.read\_csv("characters.csv", sep=",")

starplanets = pd.read\_csv("planets.csv", sep=",")

starspecies = pd.read\_csv("species.csv", sep=",")

starships = pd.read\_csv("starships.csv", sep=",")

starships = pd.read\_csv("starships.csv", sep=",")

starvehicles = pd.read\_csv("vehicles.csv", sep=",")

stars = starchars[["species","name"]].groupby("species").count()

stars = stars.sort\_values(by="name", inplace=False ,ascending=False)

stars.head(1)

starsp = starchars[["homeworld","name"]].groupby("homeworld").count()

starsp = starsp.sort\_values(by="name", inplace=False ,ascending=False)

starsp = starsp.head(2)

starsp

starship = starships.sort\_values(by="cost\_in\_credits", inplace=False ,ascending=False)

starship.head(1)

Q2.PySpark

1. What is an RDD?

An RDD is a Resilient Distributed Database which is immutable meaning no changes can be made to it.

1. What is a DAG?

A Dag stands for Directed Acyclic Graph it is where the operations such as transformations and action in spark take place on spark. Once an Action is performed, the DAG then submits the job to a DAG scheduler.

1. What is the role of a spark driver in the spark cluster?

The spark driver is the one that the declares all actions and transformations that are happening on RDDs and submits it to the master. Essentially it is the driver that sets the SparkContext which allows the script to connect to the Master.

1. Is Spark fault Tolerant and how does Spark achieve that?

Spark uses data storage like HDFS which splits the data into many blocks and copies of the block sit on various data nodes and so any data lost on a node can be recovered by another data node. Since RDD are immutable datasets, they cannot be transformed so Spark has copies of the original RDDS and in case of a failure, it can recover the original RDDS before transformations and actions were performed.

Q3. Shell

How to make a shell script executable ?

You apply chmod +x <script> and then run it with ./<script> command

b. What is the use of “#!/bin/bash” ?

This tells the script to run through the bash shell and not any other shell

c. How do you resolve variable in a shell script ?

Q4. Hadoop

What are the core components of Hadoop?

What is the difference between nameNodes and dataNodes?

What does jps command do in Hadoop?

What do you mean by metadata in Hadoop?

What is a block in HDFS, why block size 64MB?