1.)

a.) Relational database
b.) Streaming engine
c.) Document-oriented store
d.) Graph database
e.) Key-Value store
f.) Column-oriented store

2.)

a.)

	A	B	X	Y
A	0	1	0	1
B	0	0	1/2	0
X	1/2	0	0	0
Y	1/2	0	1/2	0

$$\begin{matrix}0 & 1 & 0 & 1\\ 0 & 0 & .5 & 0\\ .5 & 0 & 0 & 0\\ .5 & 0 & .5 & 0\end{matrix} \times \begin{matrix}1\\2\\3\\4\end{matrix} = \begin{matrix}6\\1.5\\0.5\\24\end{matrix}$$

$$\begin{matrix}0 & 1 & 0 & 1\\ 0 & 0 & .5 & 0\\ .5 & 0 & 0 & 0\\ .5 & 0 & .5 & 0\end{matrix} \times \begin{matrix}6\\1.5\\0.5\\24\end{matrix} = \begin{matrix}3.5\\.25\\3\\3.25\end{matrix}$$

$$\begin{matrix}0 & 1 & 0 & 1\\ 0 & 0 & .5 & 0\\ .5 & 0 & 0 & 0\\ .5 & 0 & .5 & 0\end{matrix} \times \begin{matrix}3.5\\.25\\3\\3.25\end{matrix} = \begin{matrix}3\\1.5\\1.5\\3\end{matrix}$$

$$\begin{matrix}0 & 1 & 0 & 1\\ 0 & 0 & .5 & 0\\ .5 & 0 & 0 & 0\\ .5 & 0 & .5 & 0\end{matrix} \times \begin{matrix}3\\1.5\\1.5\\3\end{matrix} = \begin{matrix}4.5\\.75\\1.75\\2.5\end{matrix}$$

$$\begin{matrix}0 & 1 & 0 & 1\\ 0 & 0 & .5 & 0\\ .5 & 0 & 0 & 0\\ .5 & 0 & .5 & 0\end{matrix} \times \begin{matrix}4.5\\.75\\1.75\\2.5\end{matrix} = \begin{matrix}3.25\\.875\\2.25\\3.125\end{matrix}$$

$$\begin{matrix}0 & 1 & 0 & 1\\ 0 & 0 & .5 & 0\\ .5 & 0 & 0 & 0\\ .5 & 0 & .5 & 0\end{matrix} \times \begin{matrix}3.25\\.875\\2.25\\3.125\end{matrix} = \begin{matrix}4\\1.125\\1.625\\2.75\end{matrix}$$

Page Rankings: A = 1
Y = 2
X = 3
B = 4

b.)

	A	X	Y	Q	P
A	0	1/2	1/2	0	0
X	0	0	1/3	0	0
Y	1	0	0	0	0
Q	0	1/2	1/3	0	0
P	0	0	0	1	0

To calculate rank of Q, we must drop P.

$$\begin{matrix} 0 & 1 & .5 \\ 0 & 0 & .5 \\ 1 & 0 & 0 \end{matrix} \times \begin{matrix} .333 \\ .333 \\ .333 \end{matrix} = \begin{matrix} .5 \\ .166 \\ .333 \end{matrix}$$

$$\begin{matrix} 0 & 1 & .5 \\ 0 & 0 & .5 \\ 1 & 0 & 0 \end{matrix} \times \begin{matrix} .5 \\ .166 \\ .333 \end{matrix} = \begin{matrix} .333 \\ .166 \\ .5 \end{matrix}$$

Page rankings of is A = 2, X = 3, Y = 2

Page rank of Q = (1 * 0) + (.5 * .166) + (.333 * .5)
= 0 + . 083 + .166
= .25 or ¼

Page rank of P = 1 * ¼
= .25 or ¼

c.)

The first node contains a loop, which means that the page redirects back to itself and a node. The rank of this page would be 1. To calculate the page rank of the next node, would be 1 x ½, which results in a page rank of ½. This tells us that the page rank for each of the following dead end nodes is ½.

3.)

a.)

(6 + 16 + 17 + 28 + 10 + 20 + 21 + 22 + 23 + 28 + 26 + 30) / 12 = $20.58

b.)

1pm – 4pm = (6 + 16 + 17) / 3 = 13
4pm – 7pm = (28 + 10 + 20) / 3 = 19.333
7pm – 10pm = (21 + 22 + 23) / 3 = 22
10pm – 12am = (28 + 26 + 30) / 3 = 28

c.)

1pm – 4pm = (6 + 16 + 17) / 3 = 13
3pm – 6pm = (17 + 28 + 10) / 3 = 18.33
5pm – 8pm = (10 + 20 + 21) / 3 = 17
7pm – 10pm = (21 + 22 + 23) / 3 = 22
9pm – 12am = (23 + 28 + 26) / 3 = 25.666
11pm – 2am = 0 because there are only hours from 11 and 12, no hours from 1am and 2am.

4.)

Mapper_1.py :

```
#!/usr/bin/python
import sys

# lo_quantity = field 8
# lo_revenue  = field 12
# lo_discount = field 11
# lo_orderpriority = field 6

for line in sys.stdin:

   line = line.strip()
   vals = line.split("|")

   if vals[6].endswith('URGENT'):
       lo_rev = vals[12]
       lo_qua = min(vals[8])
       lo_dis = min(vals[11])

       print '%s\t%s\t%s' % (lo_rev,lo_qua,lo_dis)
```

Mapper_2.py :

```
#!/usr/bin/python

import sys

# lo_rev = field 0
# lo_qua = field 1
# lo_dis = field 2

for line in sys.stdin:

    line = line.strip()
    vals = line.split('\t')

#     print (vals[0])
    revenue  = vals[0]
    quantity = vals[1]
    discount = vals[2]

    if ( int(discount) > 6 and int(discount) < 8):
        print '%s\t%s' % (quantity,revenue)
```

map1_map2_reducer.py:

```
#!/usr/bin/python

import sys

curr_id = None
id      = None
cnt     = 0

for line in sys.stdin:
```

```
line = line.strip()
    quan, rev = line.split('\t')

    if curr_id == id:
        cnt += 1
    else:
        if curr_id:
            print( '%s\t%s' % (quan, rev)  )
        curr_id = id
        cnt = 1

if curr_id == id:
    print ( '%s\t%s' % (quan, rev))
```

5.)

a.)

```
# Stanford web graph from 2002
# Nodes: 281903 Edges: 2312497
```

b.) & c.) Unfortunately, after executing the "time hadoop jar PageRank….." command, the execution seems to hang. After 30 minutes, I stopped the first execution and started a new job. Again, the results were the same, the execution seems to hang.

```
/home/ec2-user/hadoop-pagerank/src
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$ hadoop fs -mkdir /data/
mkdir: `/data': File exists
[ec2-user@ip-172-31-11-74 src]$ hadoop fs -mkdir /data/webStanford
[ec2-user@ip-172-31-11-74 src]$ hadoop fs -put ~/web-Stanford.txt /data/webStanford
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$ hadoop fs -ls /data/webStanford
Found 1 items
-rw-r--r--   2 ec2-user supergroup   32888333 2022-05-30 17:21 /data/webStanford/web-Stanford.txt
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$ time hadoop jar PageRank.jar it.uniromal.hadoop.pagerank.PageRank --input /dat
a/webStanford --output /data/prOutput --damping 90 --count 8
Damping factor: 1.0
Number of iterations: 8
Input directory: /data/webStanford
Output directory: /data/prOutput
---------------------------
Running Job#1 (graph parsing) ...
22/05/30 17:21:59 INFO client.RMProxy: Connecting to ResourceManager at /172.31.11.74:8032
22/05/30 17:22:00 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implem
ent the Tool interface and execute your application with ToolRunner to remedy this.
22/05/30 17:22:00 INFO input.FileInputFormat: Total input paths to process : 1
22/05/30 17:22:00 INFO mapreduce.JobSubmitter: number of splits:1
22/05/30 17:22:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1653931204343_0001
22/05/30 17:22:00 INFO impl.YarnClientImpl: Submitted application application_1653931204343_0001
22/05/30 17:22:01 INFO mapreduce.Job: The url to track the job: http://ip-172-31-11-74.us-east-2.compute.inter
nal:8088/proxy/application_1653931204343_0001/
22/05/30 17:22:01 INFO mapreduce.Job: Running job: job_1653931204343_0001

^C
real    46m53.514s
user    0m9.749s
sys     0m1.040s
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$
[ec2-user@ip-172-31-11-74 src]$ time hadoop jar PageRank.jar it.uniromal.hadoop.pagerank.PageRank --input /dat
a/webStanford --output /data/prOutput --damping 90 --count 8
Damping factor: 1.0
Number of iterations: 8
Input directory: /data/webStanford
Output directory: /data/prOutput
---------------------------
Running Job#1 (graph parsing) ...
22/05/30 18:10:16 INFO client.RMProxy: Connecting to ResourceManager at /172.31.11.74:8032
22/05/30 18:10:16 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implem
ent the Tool interface and execute your application with ToolRunner to remedy this.
22/05/30 18:10:16 INFO input.FileInputFormat: Total input paths to process : 1
22/05/30 18:10:16 INFO mapreduce.JobSubmitter: number of splits:1
22/05/30 18:10:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1653931204343_0002
22/05/30 18:10:17 INFO impl.YarnClientImpl: Submitted application application_1653931204343_0002
22/05/30 18:10:17 INFO mapreduce.Job: The url to track the job: http://ip-172-31-11-74.us-east-2.compute.inter
nal:8088/proxy/application_1653931204343_0002/
22/05/30 18:10:17 INFO mapreduce.Job: Running job: job_1653931204343_0002
```