Aufgabe (10 Punkte)

a) Was versteht man unter dem CAP-Theorem? Erklären Sie auch kurz was ein „Theorem“ ist. (5 Punkte)
b) Diskutieren Sie dazu folgende Aussage: (5 Punkte)

„*Zu behaupten, ein Datenhaltungssystem sei konsistent und verfügbar, nicht aber partitionstolerant, ergibt überhaupt keinen Sinn.“*

Aufgabe (11 Punkte)

a) Beschreiben Sie stichwortartig das Map/Reduce Verfahren oder definieren Sie es per Formel. (3 Punkte)
b) Wenden Sie anschließend das Verfahren Map/Reduce auf die untere Datentabelle (Passagieraufkommen Januar 2017 bis Dezember 2017) an, indem Sie das Map/Reduce Verfahren Schritt für Schritt dokumentieren (Input bis Final Result). Beantworten Sie über dieses Verfahren die Frage, welche Flugverbindung* die meisten Passagiere im Zeitraum Januar 2017 bis einschließlich Dezember 2017 transportiert hat. Wieviel Passagiere ware es genau? (8 Punkte)

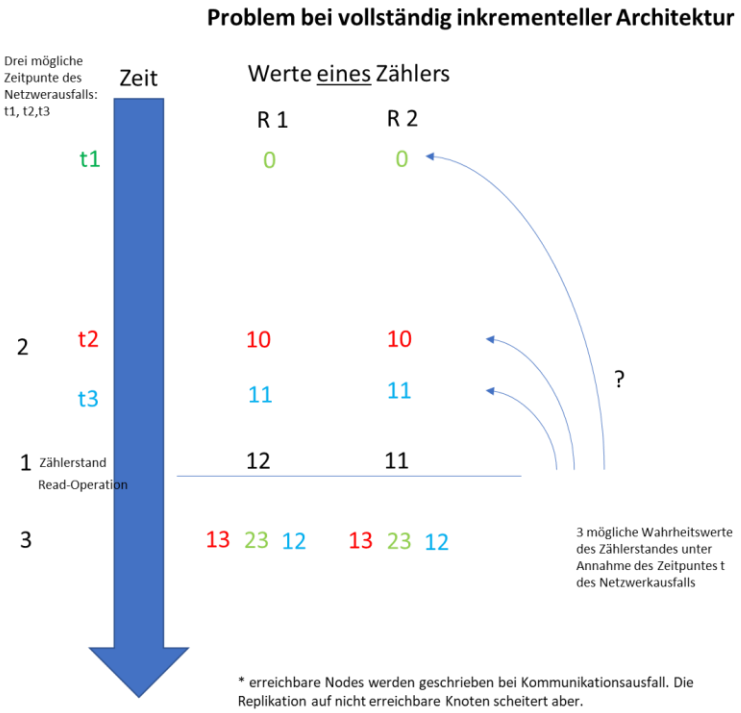
*Eine Flugverbindung bedeutet gleicher Abflug-Flughafen und gleicher Ankunft-Flughafen. Aufgrund unterschiedlicher Ablug-bzw. Ankunftszeiten oder Kapazitäten kann eine Flugverbindung unterschiedliche Flug-IDs haben.

Abflug	Ankunft	Flug-ID	Anzahl Passagiere/Flug-ID
DUS	SFO	347	411
DUS	MUC	943	204
HAM	RDM	666	322
DUS	SFO	348	222
LAS	NYC	147	221
STG	LCY	369	298
AMS	LHR	258	228

Tabelle: Passagieraufkommen Januar 2017 bis Dezember 2017

Aufgabe (9 Punkte)

Betrachten Sie das aus der Vorlesung bekannte Beispiel eines Zählerstandes:



a) Erklären Sie die Graphik mithilfe der aufgeführten Schritte 1, 2 und 3 (Reihenfolge). (5 Punkte)
b) Welches Problem ergibt sich (Problembeschreibung) ? (2 Punkte)
c) Welche Lösung schlagen Sie vor? (2 Punkte)

Aufgabe (6 Punkte)

R produziert untere Ausgabe. Der Datensatz tips sollte aus der Vorlesung bekannt sein.

```
favstats (total_bill ~ 1, data = tips)
```

```
##     1  min    Q1    median    Q3   max    mean       sd   n missing
## 1   1 3.07 13.3475  17.795 24.1275 50.81 19.78594 8.902412 244       0
```

a) Wie berechnen Sie die Quadratische Abweichung aller Rechnungen aus der obigen Ausgabe? (4 Punkte)
b) Berechnen Sie die Quadratische Abweichung aus den obigen Daten. (2 Punkte)

Aufgabe (9 Punkte)

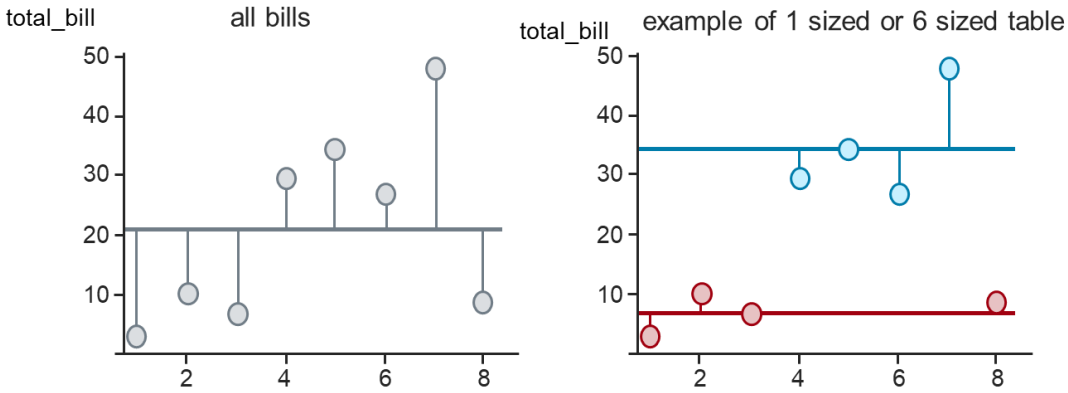
Der Befehl favstats unter R wird im Folgenden modifiziert (im Vergleich zu Aufgabe 3):

```
favstats (total_bill ~ size, data = tips)
```

R erzeugt dazu folgende Ausgabe:

```
##   size  min     Q1  median    Q3   max    mean       sd   n missing
## 1    1 3.07 6.2050   7.915 8.9525 10.07 7.24250 3.010729   4       0
## 2    2 5.75 12.4525 15.370 19.6900 40.55 16.44801 6.043729 156       0
## 3    3 10.33 16.9400 20.365 27.7750 50.81 23.27763 9.407065  38       0
## 4    4 16.49 21.5000 25.890 34.8100 48.33 28.61351 8.608603  37       0
## 5    5 20.69 28.1500 29.850 30.4600 41.19 30.06800 7.340396   5       0
## 6    6 27.05 29.1125 32.050 37.7675 48.17 34.83000 9.382000   4       0
```

Die graphische Ausgabe dazu ist hier angegeben. Das rechte Bild zeigt das Beispiel von 1-Person und 6-Personen Tischen.




```
favstats (total_bill ~ 1, data = tips)
```

```
favstats (total_bill ~ size, data = tips)
```

a) Was ist der Unterschied in der Bedeutung der Befehle (siehe Aufgabe 3 and 4)? (2 Punkte)
b) Berechnen Sie die Quadratische Abweichung der Tische mit 1-Person Größe und 6-Personen Größe. (3 Punkte)
c) Interpretieren Sie die graphische Ausgabe. (3 Punkte)

d) Was denken Sie: Ist die Summe der Quadratischen Abweichung über alle Rechnungen mit Tischgrößen-Kategorisierung höher oder niedriger als ohne Tischgrößen-Kategorisierung?
Begründen Sie. (1 Punkt)

Aufgabe (7 Punkte)

Er wird folgende Ausgabe bei einer lineraren Regression unter R erzeugt:

```
## Call:
## lm(formula = tip ~ total_bill, data = tips)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1982 -0.5652 -0.0974 0.4863 3.7434
##
## Coefficients:
##             Estimate Std.    Error t value Pr(>|t|)
## (Intercept)  0.920270  0.159735   5.761  2.53e-08  ***
## total_bill   0.105025  0.007365  14.260   < 2e-16  ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 242 degrees of freedom
## Multiple R-squared: 0.4566, Adjusted R-squared: 0.4544
## F-statistic: 203.4 on 1 and 242 DF, p-value: < 2.2e-16
```

a) Wie lautet die geschätzte Gleichung für die Tringeldhöhe $\widehat{tip}$? (3 Punkte)
b) Wie lautet die Prognose für das Trinkgeld bei einer Rechnungshöhe (total_bill) von 10$? (2 Punkte)
c) Für welchen Fall stimmt die Aussage, dass das prognostizierte Trinkgeld mit dem tatsächlich gezahlten Trinkgeld übereinstimmt? (2 Punkte)

Aufgabe (5 Punkte)

a) Erklären Sie das untere R Programm Zeile für Zeile . Die Code-Zeilen sollten aus der Vorlesung bekannt sein. (3 Punkte)

```
set.seed(1896)
Sarah_Raet <- do(1000) * rflip(n = 8)
prop( ~ heads, success = 8, data = Sarah_Raet)
```

b) Erklären Sie untere Ausgabe in R? (2 Punkte)

```
## prop_8
##  0.001
```