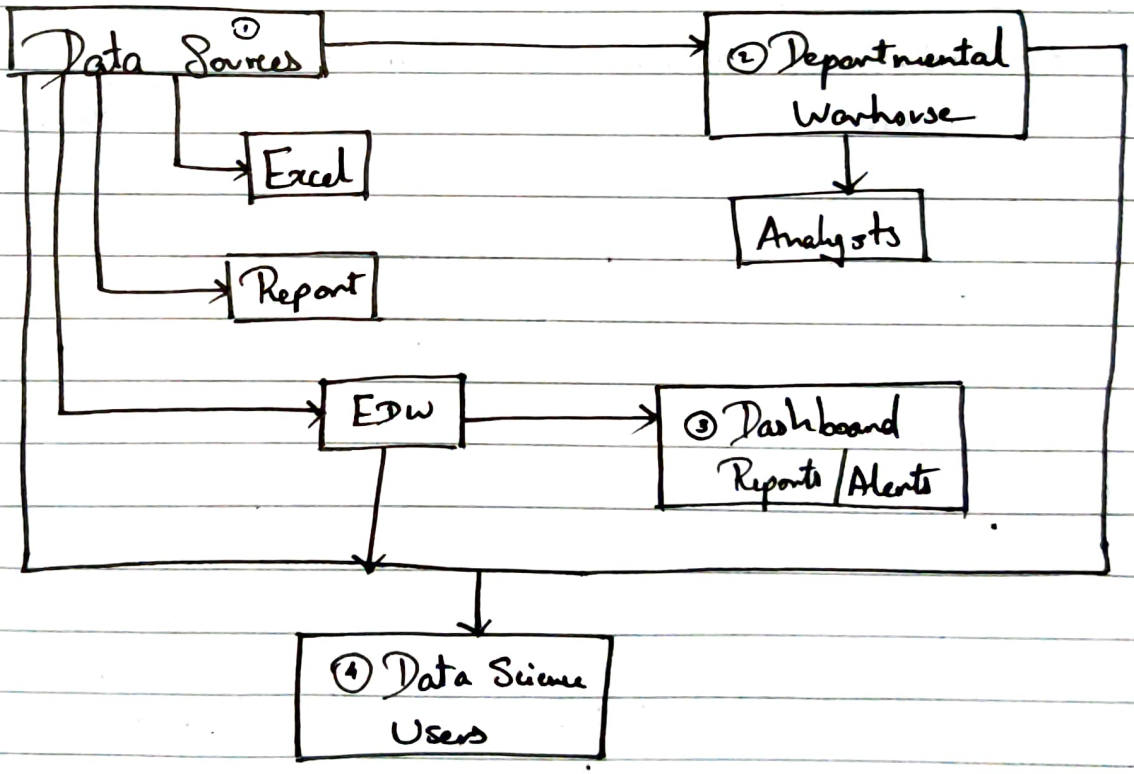
* 26/08/2020
* Data Analytics
Shiva Saran
BE B - 20

* Unit Test 1

Question 1

* Current Analytical Architecture



* ~~Anaytic~~ Analytic Architecture.

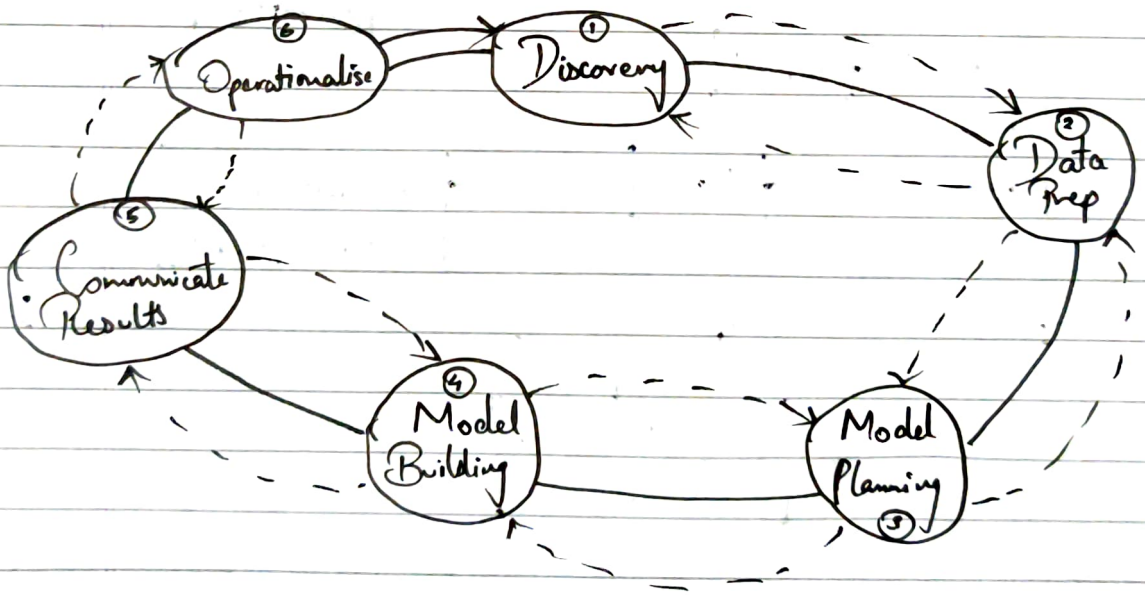
① Data Sources
For the data to be loaded into the data warehouse, there is a need to normalise the data with suitable data type definitions & in a structed format.

② Departmental Warehouse
Data is read by additional applications across the enterprise for BI & reporting purposes.

③ EDW
It achieves the objective of reporting & creation of dashboard. It generally limit the ability of analyst to iterate on the data in a separate non-prod enviroment where they conduct in-dept analysis.

Question 2

* Data analytical life cycle:
It has 6- phases as mentioned below:



Phase 1: Learn the business domain & its related data.
Phase 2: Presence of analytic sandbox.
Phase 3: Determine the models, techniques & workflow it intends to follow.
Phase 4: Develop data sets for testing, training & production

Phase 5: In collaboration with major stakeholders, determine the results of the project.
Phase 6: Deliver final reports, briefing code & technical documents & run the pilot project.

Question 3

(a) Wilcoxon Rank Sum test
It is a non-parametric hypothesis test that checks whether two populations are identical distributed. Wilcoxon test does not assume anything about the population distribution, it is generally considered more robust than the t-test. This test is often performed as a two-sided & thus the research hypothesis indicates that the populations are not equal as opposed to specifying directionality.

(b) Type 1 & Type 2 Error

- Type 1 error
A type 1 error is also known as false positive & occurs when a researcher incorrectly rejects a true null hypothesis. This means that your report that your findings are significant when in fact they have occured by chance.

- Type 2 Error
A Type 2 error is also known as false negative and occurs when a researcher fails to reject a null hypothesis which is really false.

(c) ANOVA

Analysis of variance (ANOVA) is generalisation of hypothesis of the difference of 2 population means. ANOVA test if any of the population means differ from the other population means. The null hypothesis is that all the population means are equal.

Question 4

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫
Height	185	170	168	179	182	188	180	180	183	180	180	177
Weight	72	56	60	68	72	77	71	70	84	88	67	76

→ Initial K1 = 185 & 72

Initial K2 = 170 & 56

Euclidean Distance for	K1	K2	New Centroid K1	New Centroid K2
③	20.81	4.47	185 , 72	169 , 58
④	7.21	14.14	182 , 70	169 , 58
⑤	2.00	19.10	182 , 71	169 , 58
⑥	8.49	26.87	185 , 74	169 , 58
⑦	5.83	17.03	183 , 73	169 , 58
⑧	3.54	16.28	181 , 71	169 , 58
⑨	12.87	29.53	182 , 78	169 , 56
⑩	10.59	31.95	181 , 83	169 , 58
⑪	15.85	14.21	181 , 83	175 , 63
⑫	8.06	13.73	179 , 80	175 , 63

Grouping:

K1 → ① , ④ , ⑤ , ⑥ , ⑦ , ⑧ , ⑩ , ⑫

K2 → ② , ③ , ⑪

* Question 5

Regression is statistical method used in finance, investing and other disciplines that attempts to determine the strength & character of the relationship between one dependent value & a series of other variables.

There are 2 basic types of regression:

① Simple linear regression
② Multiple linear regression.

Linear regression establishes a relationship between dependent variable (Y) & one or more independent variables (X) using a best-fit straight line.
It is represented by an equation $Y = a + b * X + e$,
where: a is interept
b is slope of the line
e is error term.