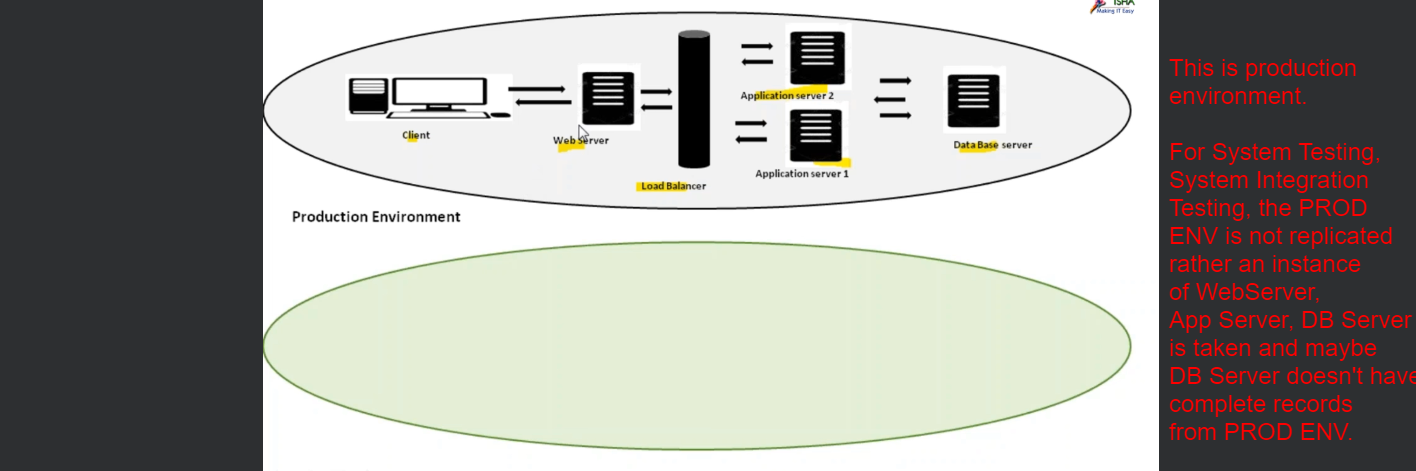
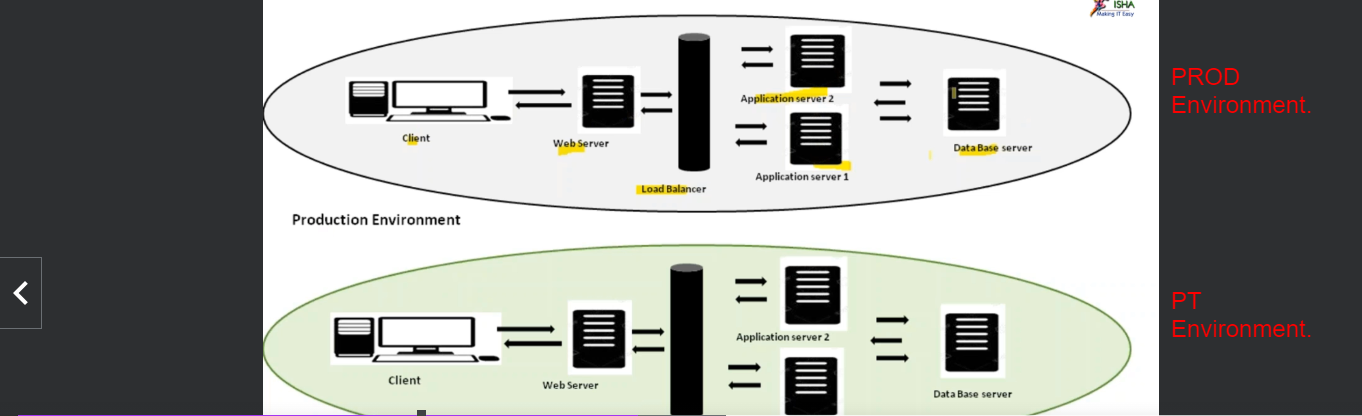
1. Let’s try to understand scalability.
2. Let’s take Mark’s example again to understand scalability.
3. Mark already tested BMW car for speed (for one hour, the car speed is fine with current peak load), stability (for 8 hours with less than current peak load, the car’s speed is fine).
4. Mark wants to test his car so that it works fine in future as he is planning to have 3 kids.  
     
   So, Mark wants to make sure that with that future load, the car will drive fast enough.
5. He took 3 kids from some family and now he is driving his car with 8 people (5 already from his family + 3 kids) to test if the car will drive fast enough **with future load**.  
   This is called **testing for scalability**.

5 People here is **current load**. And when 8 people, it is **future load**.

1. In our world (PT), when we tested for speed, we tested for 1000 users but we know for this Gmail app, probably six months from now, it will not be 1000 users, but there will be more people using this Gmail as Gmail is very popular Email.
2. So, as a performance tester, we not only test the app for the **current load** but also for **future load**.
3. So, performance testing is for
   1. **Speed**: With current peak load, app is responding under SLA.
   2. **Stability**: With normal load, the app is responding under SLA and without breakdown.
   3. **Scalability**: The app is responding under SLA with future load (which is more than current Peak Load).
4. What is **Production Like Environment**?
5.   
   The same is true for DEV Environment.
6. But for our performance environment, we will compromise with the environment setup.   
   If they give instances, we will reject.  
   But we want is a production like environment.  
     
   So, in production if we have one web server, 2 app servers and 1 DB server, same we want in PT Environment and only then we call it “Production Like Environment”.  
   Because the response time may get affected based on how many web servers, app servers, DB Servers we have in PT Environment.  
   If Hardware configuration is good, then most probably the response time is good.  
   Not only hardware configuration but software configuration should be same in PT environment.  
   Even RAM, Hard-Disk, Processor Speed, no of records should be same.   
   only then we as Performance Tester can confirm if the app will respond fast or slow in production.  
   That is why ideally, both the environments (PROD, PT) should be similar.   
   “Ideally” because sometimes it is not ideal to have same configurations as it requires enough time and it also costs a lot as a lot of different category people are involved such as Network Guys, the Server Guys, DB Server Guys, the DEVELOPER etc.  
   But some clients go through that pain as they want PT to be done correctly.
7. This is the case all the time?  
   No, sometimes we use **Scaled Down Environment**.   
   Let’s say you have two app servers but we want 1 app server.  
   This is when you have to set the expectation to the client that whatever the response time you get as part of performance testing team, it might or might not be the same in production environment.
8. But as per the PT definition, PT should be done on “Production Like Environment”.
9. 